Contextual Computing for Natural Language Processing

Robert Porzel

Contents

List of Tables

List of Figures

Acknowledgments

First and foremost I want to thank my family and friends, especially Carolin, Simon and Markus for their help, patience and constant support as well as Dieter and Käte for getting everything started in the first place and helping out whenever needed. Next, I want to thank and acknowledge numerous colleagues, collaborators as well as founding sources, which I will do in reference to their specific and kind contribution to this work. First and foremost I want to thank Professors Rainer Malaka and Jerry Feldman for their excellent supervision and advice. Additionally, Professors Wolfgang Wahlster, Andreas Reuter, Peter Hellwig, Ralf Klabunde and Klaus Tschira for their efforts in linking the individual research projects to a national and international consortium of interdisciplinary partners and for providing very direct input and help to me. Through them and the German Research Foundation support the SFB 245 *Language and Situation*, the SPP 1022 *Language Production* were made possible. Through the Klaus Tschira Foundation grants for the *DeepMap* and *EDU* projects as well as the KTS co-founding of the *SmartKom* and *SmartWeb* projects - that were also founded by the German Ministry of Research and Education (BmbF) - this work was made possible. Now, for the specific advice, help and contributions on the work presented in Chapter 3, I want to thank my colleagues at the time, especially Iryna Gurevych, Rainer Malaka and Michael Strube. For their contribution to the annotation experiments and implementations I want to thank numerous students, especially Hans-Peter Zorn for Section 3.2, Christof Müller for Section 3.3, Berenike Loos for Section 3.3 and Vanessa Micelli for Section 3.5.

Also, for their help and contributions on the work presented in Chapter 4, I want to thank Ralf Klabunde and Daniel Glatz for their input for Section 4.1.1 and my EDU and SmartWeb colleagues at the University of California in Berkeley and the University of Bremen and the European Media Laboratory, i.e. Rainer Malaka, Jerry Feldman, Nancy Chang, Ben Bergen, John*o* Bryant, and, of course, the Heidelberg *SmartCoffee* Gang (Hidir Aras, Jürgen Vogel, Hans-Peter Zorn, Berenike Loos and Vanessa Micelli) for their input in Section 4.2 as well as for their advice and great collaboration on many issues. Within the DeepMap, EDU, SmartKom and SmartWeb projects, this work has been performed in concert with numerous other collaborators who where working at ICSI, DFKI, CMU as well as at the universities of Berkeley, Bremen, Heidelberg, Karlsruhe and Munich at the time. I would like to thank them for the ensuing cooperations and publications on mutual areas of interest and the plea-

sure of meeting them. Last, but not least, I want to express my gratitude to the many researchers that provided prior work cited and - as in the case of Charles Fillmore's or Aldo Gangemi's work - employed herein.

Chapter 1

Introduction

> Words exist because of meaning; once you have grasped the meaning, you can forget the words. Where can I find a man who has forgotten words, so I can have a word with him? [Chuang, 300] Chapter:26

One of the goals in artificial intelligence concerns the creation of intuitively usable interfaces that lower the gap between increasingly complex applications and their users. Consequently, sub-fields have emerged that seek to develop intelligent user interfaces as well as ubiquitous and pervasive computing environments for intuitive *everyday* computing. For this, the notion of *hands-free computing* has received considerable attention, especially due to the advent of mobile computing, which has been pushed forward due to increasingly small and powerful devices for which traditional interaction paradigms are more or less unsuitable [Johnson, 1998].

Therefore, a new window of opportunity for spoken dialog systems opened through which research on human-computer interaction via natural language was moved away from artificially reduced blocks-world scenarios to realistic applications and into the hands of mobile users. Today, so called *controlled* dialog systems [Allen et al., 2001a] have become reliable enough to be deployed in various real world applications, e.g. timetable or cinema information systems. In controlled dialog systems the interaction between the user and the system is restricted so that recognition and understanding errors can be kept minimal. The more *conversational* a dialog system becomes, the less predictable are the users' utterances. Recognition and processing, therefore, become increasingly difficult and unreliable. This is due to the fact that virtually on all levels in the natural language processing pipeline, ambiguities, underspecification as well as *noise* multiply greatly.

An important step toward the development of more natural and intuitively usable dialog systems was constituted by the inclusion of additional modalities such as gesture, mimics or haptics. One of the central motivations behind the work on multi-modal systems is based on the hypothesis that the individual modalities can be employed to mutually disambiguate each other.

A single pointing gesture, for example, can be as ambiguous with regards to the user's intention as the utterance *tell me more on about that.* Fused together, however, they provide enough information for a system to infer the intended action requested. In recent years several research projects on multimodal interfaces sought to overcome individual problems arising with more or truly conversational dialog systems [Malaka and Zipf, 2000, Allen et al., 2001b, Wahlster et al., 2001, Johnston et al., 2002, Sonntag et al., 2007]. The goal of intuitive and conversational multimodal interfaces that can someday be used in real world applications, however, has not been achieved yet.

The work presented herein is to be understood as part of the larger research undertaking pursuing this goal of intuitively usable multimodal systems that can cope robustly with spoken language and other modality-specific input. For this, I will examine the following research question and hypothesis: While it is true that individual modalities disambiguate each prior approaches have by and large overlooked - or at least failed to treat - contextual information as another modality that is crucial for interpreting the standard modality-specific input, such as user utterances, gestures and the like, felicitously. In short, the claim is, that systems seeking to understand natural and conversational input context needs to be treated as a *bona fide* modality that contributes equally pertinent information as all other modalities and, therefore, needs to be considered analogously. Failing to include context consequently causes systems to become restricted, brittle and inherently unscalable.

In order to support this claim, the challange of modeling and applying of contextual - and therefore linguistically implicit - information and the corresponding pragmatic knowledge will be examined as one of major challenges for understanding conversational utterances in unrestricted dialog systems. In order to provide a more concrete motivation, I will exemplify three problems, which still thwart unrestricted conversational interaction via natural language in the following section.

1.1 Motivation

What - a frustrated computer scientist or application developer might ask - is the *raison d'être* - for ambiguities, underspecification or noise. And why are we not to live in a world where input comes in noiseless, unambiguous and clearly specified packages. Especially, looking at spoken or written natural language utterances as the key input modality for multimodal systems, one finds that they are frequently ambiguous, highly underspecified and, additionally, come with various kinds of noise. In contrast to human languages, computer languages are designed to avoid morpho-syntactic, semantic or pragmatic ambiguities. Human languages, however, seem to be riddled with situations where the addressee has to choose between multiple interpretations. In these cases linguist say that the addressee *resolves* the ambiguity. For human beings the process of resolution is often unconscious, to the point that it is sometimes difficult to recognize that there ever was any ambiguity.

Classic examples of ambiguity commonly present lexical ambiguities, e.g. for the lexical item *bank*. Some examples that can be found in the British National Corpus [BNC, 2008] are given with their respective BNC indices in parenthesis below:

(1) *... the gene bank might ultimately contain 500 collections ...* (B76 493)
(2) *Robbers ran from a bank in Milan, Italy, ...*(CBE 2870)
(3) *... south of the town, on the west bank of the River Kent ...* (B0A 1365)

Please note, that in these examples sufficient 'contextual information is provided by the lexical neighborhood to enable the reader to resolve the ambiguity, i.e. to know what was meant. This is usually stated by saying that a lexical item can have more than one *meaning* or that it exhibits different - but related - *senses*. Based on that distinction, linguists differentiate between *homonymy* in the former and *polysemy* in the latter case [Kilgarriff, 1993]. In the fields of linguistics and computational linguistics it is well known that resolving such ambiguities, i.e. finding the meaning or sense, that is at hand in a given utterance or sentence, can only be done by *observing* at least the lexical item in that particular discourse context [Ide and Veronis, 1998, Schütze, 1998, Widdows, 2003a].

In the literature this type of context is usually equated with the surrounding words and corresponding classifiers can be trained using supervised machine learning techniques [Mitchell, 1997, Stevenson and Wilks, 2001]. It is important to remember, however, that successful contextual computing in this regard does not only encompass the contextual information about the surrounding lexical items but also the contextual knowledge of the real-world relations between the entities denoted by these items, i.e. between genes and banks, rivers and banks or robbers and banks. In the supervised learning approaches this knowledge is provided by the human annotators of the training data, who can correctly resolve the ambiguities of *bank* in Examples 1 through 3 and can, therefore, annotate the individual occurrences correspondingly.

Underspecification occurs, whenever omissions are found, which are recoverable by recourse to context. For example, when asking for directions one might consider information about the source, i.e. from which place, and the goal, i.e. to which place, to be relevant to an instructional request. However, in many cases people asking for directions to places omit to specify the source, for example, as in:

(4) How do I get to the supermarket? (G25 74)

The more or less of obvious reason is that a default source, i.e. *from here*, is contextually given - or can be assumed - as is the case in most situations where people ask for directions. It is therefore not necessary for a speaker to explicate this bit of information, as one can rely on the addressee's capability to make the correct inference by recourse to a shared context. Such inferences are also computationally cheap for humans and take significantly less time than the laborious pronunciation or typing of the words *from here*.

In the field of linguistics expressions which are inherently context-dependent, e.g. anaphora, such as the pronouns *he* or *it* or the demonstratives *that* or *there* as well as the deictic expressions *here* or *now* are called *indexicals* [Bunt, 2000]. Utterances containing indexicals are - by virtue of the pervasiveness of contextual knowledge - the norm in discourse, with linguistic estimations of declarative non-indexical utterances around 10% [Barr-Hillel, 1954]. Without contextual awareness and the respective knowledge, utterances, or fragments thereof, become susceptible of interpretation in more than one way.

In contrast to the challenges presented by ambiguities and underspecification, problems with noise arise mainly due to problems in processing. In some cases, however, e.g. spoken or typed input, we also find *noise* as a result of hesitations, false starts and errors performed by the user or as a result of environmental conditions, e.g. real noise. Nevertheless, the more problematic and frequent sources of noise, in a technical sense, are produced by the modality-specific recognizers. Consider the following best and second best speech recognition hypotheses that were produced based on the utterance *I am on the Philosopher's Walk* spoken by a user to the SmartKom system to denote the starting point of a navigational request [Wahlster et al., 2001, Gurevych and Porzel, 2003]:

(5) Feature films on the Philosopher's Walk
(6) Am on the Philosopher's Walk

While both hypotheses contain some noise - due to recognition mistakes - it is clear that - in the given context - the intended meaning is more recoverable in Example 6 than in Example 5. Noise can, therefore, be generated by current input processing systems, such as automatic speech recognition, or by the user, for example, as typographical errors in text-based input. In the light of the dire need to address the problems of ambiguity, underspecification and noise in dialog systems with robust and scalable means, I will now sketch out the aim and intended contribution of this work in the following section.

1.2 Thesis Aim and Contribution

As one knows from personal experience, human-human communication works extremely well despite all of the challenges presented above, i.e. we can understand each other despite all of the ambiguities and underspecifications present in our utterances even at noisy cocktail parties. The amazing robustness of human-human communication is - at least in part - a result of our context-awareness and our corresponding pragmatic knowledge, both of which enable us to disambiguate and decontextualize our interlocutor's utterances robustly even under noisy conditions.

The work presented herein builds upon the recognition of the fact that computational approaches to any of the three aforementioned challenges can benefit from the inclusion of contextual information, real world knowledge and correspondingly reified contextual knowledge in order to recover the user's intent

from a given conversational input. The main aim, therefore, is to present a formal approach for explicating contextual information and pragmatic knowledge, that can be applied, employed and evaluated in natural language understanding systems.

The ancillary contributions I aim to make through this work are:

- a hitherto missing clear distinction between contextual information and the associated pragmatic knowledge - needed for this contextual computing approach;
- a set of applications of contextual computing for the various steps of processing the user's natural language input in dialog systems, such as hypothesis verification in automatic speech recognition, word sense disambiguation and relation extraction in semantic interpretation as well as intention recognition in contextual interpretation[1];
- evaluations of the contribution of contextual computing in these areas with regard to their performance as well as the corresponding methodological challangens, e.g. regarding measuring the performance of such adaptive systems as a whole;
- a descriptive ontology-based approach for enabling context-adaptive decontextualization of these interpretations applied in a real time multimodal prototype system.

It will, therefore, be shown herein that explicit formal knowledge models and means to observe a given context can be employed to build more scalable systems that are better equipped to handle context-dependent ambiguous, underspecified and noisy input. The central focus of this work, therefore, lies on the development of robust and scalable systems that can interact with human users using natural modalities, such as spoken language, which have evolved to facilitate efficient communication among interlocutors who share vast and rich amounts of background knowledge and which is always situated in given context. In order to situate this work in relation to these contributions, an overview of its organization will now be given in the following section.

1.3 Thesis Organization

The following work is presented in four additional chapters: State of the Art (Chapter 2), Domain and Discourse (Chapter 3), User and Situation (Chapter 4) and Conclusion (Chapter 5).

Chapter 2 provides an overview of prior work on knowledge modeling and on developing human-computer interfaces that seek to handle natural language input with an emphasis on their context-awareness and ability to perform context-dependent analyses. This section will conclude with a list of four context types,

[1] In the naming of these three processing steps I follow Allen's (1987) textbook nomenclature.

i.e. domain-, discourse, interlocutionary and situational context, which I regard to contribute distinct and pertinent information for contextual computing. From there on, I will examine the individual contribution of these different context types to a set of natural language processing tasks.

Chapter 3 presents a set of tasks that are specific to natural language understanding and different types of *classifications*, e.g., in seeking to resolve the *meaning* evoked a given *form* based on ontologically represented knowledge of the domain and discourse context. These tasks are examined individually throughout the chapter's sections as follows: Section 3.1 introduces the *ground* ontological domain representations and explicates the modeling principles employed for the specific representation chosen as well as presenting the ontology itself in subsections 3.1.1 and the following. Section 3.2 starts the subsequent empirical employment of this context-dependent domain- and discourse knowledge for the tasks of hypotheses verification together with its evaluation and results presented in subsections 3.2.1 and the following. Section 3.3 then adds discourse context via an ontological discourse model with its evaluation and corresponding results displayed in subsections 3.3.1 and the following. Section 3.3 takes on the task of word sense disambiguation with its evaluation and corresponding results given in subsections 3.4.1 and the following. Section 3.5 concludes this set of tasks with semantic relation extraction, with its evaluation and corresponding results given in subsections 3.5.1 and the following. Section 3.6 presents an overall evaluation of the domain representations employed as shown in subsections 3.6.1 and the following. Section 3.7 concludes this chapter with a summary of the empirical data and the corresponding experimental results.

Chapter 4 turns to the remaining interlocutionary and situational context, which will be presented in Section 4.1. Section 4.1.1 takes a look at the user as a source of contextual information and shows how the corresponding interlocutionary context has been examined and employed for contextual computing. Section 4.1.2 does the same for situational context and completes the types of contexts presented. Section 4.2 consequently examines the resolution of the intention behind a given utterance, i.e. asking what the illocutionary `function` behind a given - otherwise disambiguated - form-meaning pairing is and evaluated as a showcase approach for computational pragmatics. Finalizing this approach in contextual computing, leads to the presentation of the descriptive ontological model of what matters pragmatically in a given context in Section 4.3. I will, then, conclude this chapter by describing the implementation and application of the resulting model for pragmatic knowledge and the corresponding ontological design pattern in Section 4.4.

Chapter 5 concludes this work by summarizing the demonstrated results and our approach to contextual computing for natural language processing. Based on these results a reconsideration of our initial and applied categorizations and hypotheses will be cast in the light of further experimental steps and for future novel approaches that, given the differentia and models, we can examine appropriate learning techniques and suitable interactive models and behaviors. Finally, the references cited in this work and appendices describing additional details of the models, data structures and algorithms employed herein are given.

Chapter 2

State of the Art

Many areas of artificial intelligence (AI) have had to struggle in various ways in which contextual dependencies arise, e.g. knowledge representation, natural language processing or expert systems to name a few. As the importance of context is frequently glossed over in the literature, researchers noted already in the early 80s that the denotation of the term has become murkier throughout its extensions in different fields of AI, calling it a *conceptual garbage can* [Clark and Carlson, 1981]. A classic AI example of contextual computing showed how the medicinal expert system MYCIN [Wallis and Shortliffe, 1982] can benefit from contextual considerations when prescribing treatments, resulting in fewer fatal intoxications as a result from the prescription [McCarthy, 1984]. This case constitutes a classic example as it establishes a blueprint for the so-called *representational* approaches to contextual computing in AI [Dourish, 2001, Dey, 2001]. Initially, we find an expert system that prescribes treatments based solely on the diagnosed disease. The subsequent contextual addition is - as will be discussed in greater detail below - twofold: firstly, a set of parameters is defined, i.e. which other medications the user is also taking - and secondly, a set of rules stating what to prescribe if certain parameter settings hold.

The following sections I will provide an initial discussion of context definitions and the role contextual computing has played in the targeted area of artificial intelligence, natural language processing and multimodal systems.

2.1 Defining Context

A general feature of context and contextual computing is lack of consensus concerning of the word itself. Over the past years many researches in computer science and other areas provided a vast and diverse number of definitions. This has prompted researchers to employ latent semantic analysis (LSA) and principal component analysis on a corpus of 150 definitions for context to find prominent similarities and divergences [Foltz et al., 1998, Bazire and Brézillon, 2005].

As both LSA and the subsequent clustering showed, the definitions were very

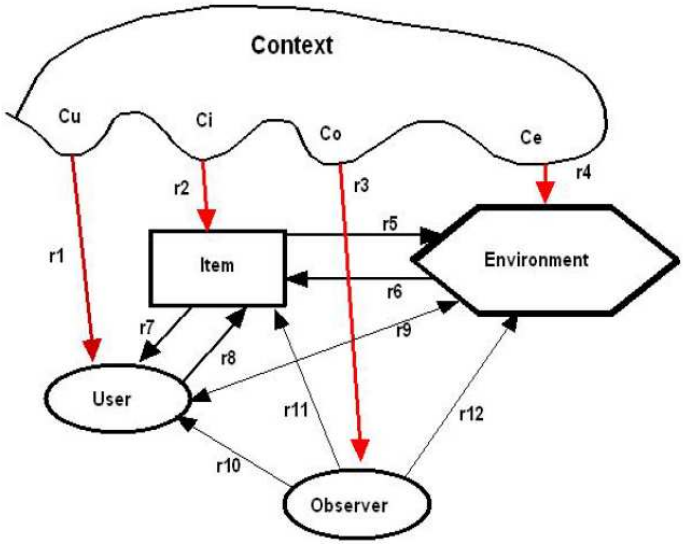

Figure 2.1: Components and relations that appear in context definitions as extracted by Bazire and Brezillon (2005)

diverse and, in general, dependent on the discipline in which they originated. Bazire and Brezillon (2005) extracted the following central components that are also shown in Figure 2.1: the user, the item, the environment, the observer and the context that influences them - as well as the relations between the context and the components and the relations among the components.

Their work shows that any definition highlights some subset of the components and relations. However, depending on the scientific area, each definition covers only a subset of the entire ensemble of components and relations and either omits or merges those components and relations that, given their own domain context, do not seem to matter. Below some sample definitions - that illustrate this domain-dependent diversity and selectivity - are listed:

- things or events related in a certain way [Ogden and Richards, 1923];
- paths of the information retrieval [Boy, 1991];
- a window on the screen [Abu-Hakima et al., 1993];
- a set of preferences or beliefs [Cahour and Karsenty, 1993];
- an infinite and partially known collection of assumptions [Turner, 1993].

In defining context for the domain of contextual computing individual definitions [Schilit et al., 1994, Dey, 2001] - as examples of the representational approach [Dourish, 2001] - constitute further instances of the model of Bazire

and Brezillon (2005). For example, Dey's definition of 2001 highlights information that can be used to characterize entities, such as person, place, or object, as well as the user and the application (components in the model shown in Figure 2.1. Analogously, from the viewpoint of context-aware computing Schilit et al (1994) highlight the location of use, the collection of nearby people and dynamically changing objects as components to which context-aware system can adapt their interaction. The critical notion of relevancy is added in Dey's definition of a context-aware system, i.e., that *it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task*. The ensuing question concerning this determination of relevancy will be discussed shortly in Section 2.2, a further terminological discussion of context in this light is also provided by Dourish (2004) from an ethnomethodological perspective [Dourish, 2004].

Given these components of the Bazire and Brezillon (2005) model and the freely specifiable relations among them, also dictionary definitions can be seen in the of components and their relations. For example, the Merriam Webster Dictionary [Merriam-Webster, 2003] defines context as: *the interrelated conditions in which something exists or occurs*. According to the Oxford English Dictionary [Soanes and Stevenson, 2005], the term *context* usually has two primary senses:

1. the words around a word, phrase, statement, etc. often used to help explain (fix) the meaning;

2. the general conditions (circumstances) in which an event, action, etc. takes place.

Clearly, the first meaning is closely related to linguistic sense and the linguists' use of the term, whereas the second sense is the one which is closer to a desirable account of context in AI. This is also congruent to the observation by McCarthy (1986) who states that:

> [A]lmost all previous discussion of context has been in connection with natural language. However, I believe the main AI uses of formalized context will not be in connection with communication but in connection with reasoning about the effects of actions directed to achieving goals. It's just that natural language examples come to mind more readily.[McCarthy, 1986]

The definition of Angeles (1981) reflects the latter desideratum expressed by McCarthy more satisfactorily, as follows:

> context (L. contexere, *to weave together*. from con, 'with', and texere, 'to weave'): The sum total of meanings (associations, ideas, assumptions, preconceptions, etc.) that (a) are intimately related to a thing, (b) provide the origins for, and (c) influence our attitudes, perspectives, judgments, and knowledge of that thing.[Angeles, 1981]

Finally, a set of useful insights are presented in Collins Cobuild English Language Dictionary [Cobuild, 1995], which lists prevalent meanings of the term as follows:

1. The context of something consists of the ideas, situations, events, or information that relate to it and make it possible to understand it fully.
2. If something is seen in context or if it is put into context, it is considered with all the factors that are related to it rather than just being considered on its own, so that it can be properly understood.
3. If a remark, statement, etc. is taken or quoted out of context, it is only considered on its own and the circumstances in which it was said are ignored. It, therefore, seems to mean something different from the meaning that was intended.

Let me refer back to the work of Bazire and Brezillon's (2005) more comprehensive analysis of context definitions summarized above at this point and conclude this section by reiterating the main points one can take home from looking at the various definitions of context:

- Given that meanings - whether one considers the meaning of a verbal or non-verbal action - always arise within (or interwoven with) a given context, it becomes clear that this meaning is lost - or harder to recover - when things are taken *out of context*; in most of the cases I will examine herein they become ambiguous or underspecified.
- Given the specific entity under scrutiny only a subset of all possible components and their relations with it are pertinent for constructing that entities meaning, which explains why specific definitions of context focus only on those components and relations they deem pertinent for that entity.

In the following I will briefly provide an overview of the consequent attempt to formalize context correspondingly.

2.2 Fleshing Out Context

The basic intuition behind explicating contextual dependencies was that any given axiomatization of a state of affairs, meanings or relations presupposes an implicit context. Any explicit context model employed in processing information should, therefore, provide the information why a particular meaning can be assigned to the information and applied to the processing. In the literature this approach has often been called *fleshing out* and is considered impossible in its maximal form:

> It is seen that for natural languages a fleshing-out strategy – converting everything into decontextualized eternal sentences – cannot be employed since one does not always have full and precise information about the relevant circumstances. [Akman and Surav, 1996]

Before examining context and contextual computing in the domain of natural language I will shortly introduce the influential notions of McCarthy on context in AI [McCarthy, 1977, McCarthy, 1986]. McCarthy (1977) states that there can never be a most general context in which all stated axioms hold and everything is meaningful. This means that whenever an axiom is written it holds true only within the implicit context assumed, and one can always find a different context in which the axiom fails. Thusly, he proposes a relativized-truth-within-a-context by stating that a given statement p **is t**rue - abbreviated as *ist* - only in a given context c, which he, consequently writes as:

$$ist(p, c)$$

This states that a formal statement, such as discussed in greater detail in Section 3.1.1, called p, holds in context c. The motivation behind this formalization lies in the increased scalability, as axioms holding in a restricted blocks-world can be *lifted* to more general contexts, e.g. for systems in which context can change. Secondly,one can define vocabularies that have context-specific meanings, as frequently found in natural language. However, while this provides the formal means to employ subsumption, or in McCarthy's terms to be able to transcend a context, it leaves open the question when to transcend and where to. Taking the viewpoint of corresponding frameworks for handling dynamic domains, e.g. situation calculus [McCarthy and Hayes, 1969], of McCarthy and Hayes one has to face the so-called *framing* problem, where - from the top-down perspective - one needs to specify when a pertinent change in the background of a frame should be evoked, because its effects the meaning of something of the foreground of the frame [Mccarthy, 1979]. In so-called *representational* approach to contextual computing, the ensuing challenge is to specify when contexts are lifted/descended or become changed in the background [Dourish, 2001].

Dourish (2001) points out that current implementations of context-dependency or context-awareness in computational systems follow an almost standardized path. Firstly, a set of possible environmental states of contextually relevant parameters are defined; then, rules are implemented that try to match sensory inputs to one of the given states during runtime.[1] Within these types of applications context-awareness is fundamentally provided by such matching processes and context itself is represented by the predefined and stored set of environmental settings.

The contributions of Dourish's work (2001) are to point out not only the difficulties of determining the appropriate settings or states of the pertinent parameters, but also that the fundamental problem of this approach to contextual computing hinges of the question of how one can pre-compile all the settings and parameters that may become pertinent in advance. In his mind it is quite impossible to define these settings and parameters based solely on past research, surveys, testing, own experience, and on the purpose of the particular system alone.

[1] This matching process commonly involves a thresholding based the measured parameters and in case of ambiguous results various mediation techniques are used in order to determine a contextual state [Dey and Mankoff, 2005].

Especially in such versatile instruments as natural language it becomes virtually impossible to predict all the possible utterances and the corresponding contextual dependencies on which their interpretation might hinge. But even in seemingly less murky waters human behavior can hardly be predicted as pointed out frequently by the example of cell phone use. It can be observed that people use their mobile phone as a watch, and although one would assume that despite the common assumption that it would be uncomfortable to pull something out of one's pocket to see what time it is, however, the number of people wearing wrist-watches has decreased. A similar development affects the use of alarm clocks. Although originally intended as a "remembering function" this property is often used instead of a conventional alarm clock, especially when traveling. Lastly the employment of Short Message Services (SMS) has greatly surprised the designers of mobile phones. Originally intended as a means to relay system-related information the capacity of one message was designed to be quite limited. Despite of this limitation and a hardly intuitive interface for entering them, SMS has become an every-day way of communication among people. In order to cope with the limitation of the message length novel abbreviations have been negotiated and completely unanticipated new writing styles have emerged, e.g. the so-called *Camel Case* sentences, such as *HowAreYou*, that are found in written messages - as SMS - where spaces between letters cost as much as the letters themselves.

The examples mentioned above show that people may use and interact with technology in unexpected ways. This reveals a fundamental problem of implementing a predefined set of settings as such approaches will inevitably not scale to cover possible interactions and behavior that will occur or evolve in future. According to Dourish the reason for this problem is that context has been approached as a representational problem by assuming the following properties of context [Dourish, 2004]:

- context is a form of information, i.e. context is seen as something that can be known, represented and encoded in software systems;
- context is delineable, i.e. it is thought to be possible to define what counts as context for a specific application in advance,
- context is stable, i.e. while context may vary from application to application, it does not vary from instance to instance of an interaction with an application;
- context and activity is separable, i.e. context is taken to describe features of the environment within which an activity takes place but the elements of the activity do not belong to context itself.

I will return to these general questions concerning representational approaches to contextual computing throughout the following sections as well as in Section 5.3, but will now shift the focus to the domain of context as it relates to natural language processing and the study of human communication.

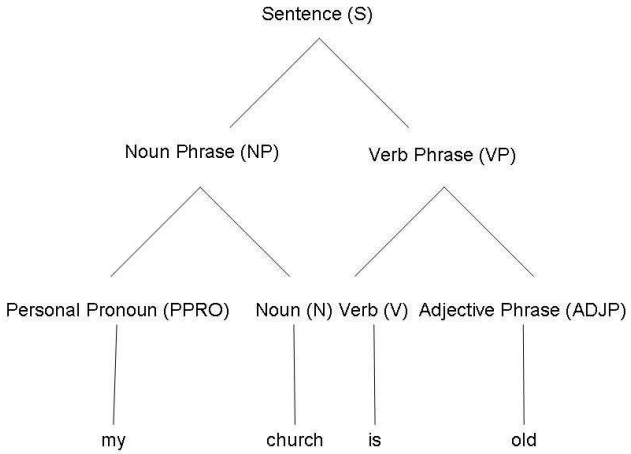

Figure 2.2: A morpho-syntactic analysis of a set of words, showing that *my* is an instance of a *personal pronoun* and *church* is one of a *noun* and both together they act as a *noun phrase*

2.3 Context in Language

In linguistics the study of the relations between linguistic phenomena and aspects of the context of language use is called *pragmatics*. Any theoretical or computational model dealing with reference resolution, e.g. anaphora- or bridging resolution, spatial- or temporal deixis or non-literal meanings requires taking the properties of the context into account. In current knowledge-based spoken dialogue systems *contextual interpretation* follows *semantic interpretation* - where the result of morpho-syntactic analysis of the automatic speech recognition (ASR) output, as depicted in Figure2.2, is mapped to logic statements [Allen, 1987].

Multimodal systems - to be discussed in greater detail below - additionally fuse the results of semantic interpretation with the results of the other modality-specific analyzers. That is, the modality-specific signals, (e.g. speech or gesture) are transferred into graphical representations (e.g. word- or gesture graphs) by means of the modality-specific recognizers, mapped onto their corresponding meaning representation and then fused using time-dependent unification techniques.

Contextual interpretation as described by Allen (1987) actually refers to the grounding of logical forms, e.g. of a logical statement expressing the proposition that my church is green. In grounding the form of a given logical statement,

e.g. the referent of the referring expression *my church* in the utterance given in Figure 2.2, a corresponding instance of the form is determined. This, however, implies that context-independent graphical and semantic representations can be computed and the context-dependent contributions follow the semantic interpretation, resulting in a final *grounded* representations. I will provide a more detailed discussion of formal models for representing the meaning of an utterance in Sections 3.1.1 and 4.1.2.

This so-called *modular* view supports a distinct study of meaning (corresponding to the semantic representation) without having to muck around in the murky waters of language use. This view is supported by the claim that some semantic constraints seem to exist independent of context. In this work,I assume a different view that also allows for context-independent constraints, but offers a less modular point of view of contextual interpretation. I will show that contextual analysis can be employed already at the level of speech recognition, during semantic interpretation and, of course, thereafter. The central claim is being made, that - as in human processing - contextual information & knowledge can be used successfully in a computational framework in all processing stages.

In recent times the so-called *modular* theory of cognition [Fodor, 1983] has been abandoned more or less completely. The so-called *new look* or modern cognitivist positions hold that nearly all cognitive processes are interconnected, and freely exchange information; e.g. influences of semantic and pragmatic features have been shown to arise already at the level of phonological processing [Bergen, 2001]. While most research in linguistics, has consequently departed from this view, most computational approaches still feature a modular pipeline architecture in that respect.

In linguistics utterances which are context-dependent are called *indexical* utterances [Bunt, 2000]. Indexical utterances are - by virtue of the pervasiveness of contextual knowledge - the norm in discourse, with linguistic estimations of declarative non-indexical utterances around 10% [Barr-Hillel, 1954]. Without contextual knowledge utterances, or fragments thereof, become susceptible of interpretation in more than one way. Computer languages are designed to avoid anaphoric, syntactic, semantic and pragmatic ambiguity, but human languages seem to be riddled with situations where the listener has to choose between multiple interpretations. In these cases one says that the listener performs *pragmatic analysis*; corresponding to contextual interpretation on the computational side. For human beings the process of resolution is often unconscious, to the point that it is sometimes difficult even to recognize that there ever was any ambiguity.[2]

The phenomenon that this process of resolution, frequently goes unnoticed is due to the fact that in many cases the ambiguity is only perceived if the contextual information & pragmatic knowledge that allowed the listener interpret the utterance unambiguously are missing. These utterances/texts, therefore,

[2]Fauconnier and Turner (2002) name some potential reasons why this may be the case from an evolutionary perspective.

become ambiguous only after they have been taken out of context, and, for example, appeared as a text(-fragment) in a linguistics textbook. The problem for computational linguistics originates - at least partially - in the fact language understanding has to make do with exactly such a contextually and pragmatically impoverished input.

Let us consider some examples and how they are treated in the literature, using the sample dialog shown in example 7.

(7) (a) User: OK, um, suppose I want to go to a museum tomorrow, which museum would you advise me?
(b) WoZ: You can visit the modern art museum.
(c) User: What is the exhibit, does it have like any architectural things inside there because I more like, you know, buildings and architectural things than, you know,

Setting up a referent is discourse is usually done by means of a referring expression (usually syntactically packaged as a noun phrase). As shown by Poesio about 50% of all noun phrases in their corpora are discourse-new, e.g. in an utterance such as shown in Example (7a) the referring expression *a museum* is considered non-anaphoric, i.e. discourse-new [Poesio and Vieira, 1998]. Anaphoric noun phrases make up 30% of their data, e.g. *it* in Example (7c) constitutes an anaphoric expression and is, hence, called the *anaphora*, which features a specific relation to its antecedent (i.e. the referring expression *the museum*). This relation is termed *co-reference* as both forms denote the same referent, i.e. a specific museum. The remaining 20% of noun phrases are made up by so-called *associative* expressions, such as bridging expressions e.g. *the exhibit* in Example (7c) is considered a bridging expression, as the employment of the definite article is licensed by the fact that the speaker assumes that the interlocutor knows that museums feature exhibits. Human annotators can reliably mark (indefinite) discourse-new and anaphoric expressions, but reliability decreases for associative expressions and those cases where discourse-new referents are introduced by definite articles, due to common world knowledge, as in *The first man on the moon* [Poesio, 2002]. The problem arises as the borderline between these cases and bridging expressions is not very clear, causing the annotator inter-reliability to decrease.

Given the distinction made by Poesio (1998) most computational approaches have focused on resolving anaphoric expressions and fewer on resolving associative expressions and handling discourse-new expressions. The most frequently studied case of anaphoric reference is that of detecting and labeling co-reference relations, where one finds a set of linguistic expressions that denote the same referent. Anaphoric expressions, however, can also range over higher level linguistic constructions in discourse, such as discussed in Byron (2002) in the case on discourse deictic expressions and abstract anaphora [Byron, 2002]. Also definite discourse-new expressions can refer to contextually evocable entities, e.g. *the old bridge* or *the mountain* in Example (8b).

(8) (a) User: and then I'd like to get out of the - out of the downtown for a

while and go on the, uh, philosopher's walk. Uh, how - how might I get there?
(b) WoZ: Um, you just, um, walk down the Haspelgasse in the opposite direction, and then you get to the old bridge. You just cross that, and then there're signs leading up to the mountain that's the philospher's walk.

Following Byron (2002) a discourse model contains so-called *discourse entities*. Discourse entities enter into the discourse model as information about events, objects, situations etc. is introduced into the discourse. Co-reference, then, means that a form part of an utterance, such as a pronoun, refers to a discourse entity that is already present in the discourse model. Setting up a referent in the discourse model, however, is not a trivial matter. In order to set up a referent correctly one has to solve various kinds of semantic and pragmatic problems that have been discussed in linguistic research falling under the categories of polysemy, metonymy, one-pronominalization, gapping and other forms of so-called *non-literal* expressions such as metaphoric expressions that will be discussed in more detail in Section 2.7.2.

However, as noted numerously in the literature natural language permits speakers to coerce terms in various ways. Coercion effects, in turn, affect pronominalization and, therefore, the resolution of anaphoric expressions. The consequence is that, unless, the discourse entities corresponding to the discourse-new expression are set-up correctly in the discourse model, anaphoric and other co-referential relations will become unresolvable by recourse to discourse context alone - for example in all cases where (grammatical) gender between the metonymic expression and the target referent differ.

Speakers can, therefore, employ extra-linguistic domain knowledge to introduce discourse-new discourse entities with definite articles, as in the case of metonymy or situational knowledge in the case of situationally-evoked referents. The same knowledge stores can be used to produce elisions and *contextual anaphora*, as in Example (9).

(9) a) User: Where is the castle?
b) WoZ: (spatial instructions)
c) User: How much does it cost?

As noted in linguistic analysis [Nunberg, 1987, Hobbs, 1991, Markert, 1999] metonymy and bridging phenomena are grounded on the fact that the given form and the referent exhibit a specific relation, called *pragmatic function* by Nunberg (1987), e.g. that museums feature exhibits licenses the bridge found in Example (7c). A bridging expression such as found in Example (10) bases on the same relation between tourist sites and their fees as the anaphora in Example (9c) exemplifies.

(10) The most popular site is the Heidelberg castle. The admission fee is 2 Euros.

For Bunt (2001) the relations between linguistic expressions and contextual settings - e.g. in the case of indexical expressions - are:

(a) expressions encoding or seeking information about aspects of contexts, e.g. about objects introduced earlier, situationally evoked referents or the relative time of speaking

(b) expressions that carry presuppositions, conversational implicatures and mappings based on shared beliefs and knowledge

In both cases the partial information encoded by the linguistic expression must be explicated relative to the given context in order for the expression to have a fully determined meaning. Expressions in which one finds (a) or (b) can thus only be understood through the relations between linguistic aspects and aspects of context. It follows that at least 90% of all declarative utterances cannot be understood if some information provided by contextual information and the corresponding knowledge is missing. In the following,I will show how the challenges have been addressed in the field of natural language processing.

2.4 Context in Natural Language Processing

Following [Allen et al., 2001b], one can differentiate between controlled and conversational dialogue systems. Since controlled and restricted interactions between the user and the system decrease recognition and understanding errors, such systems are reliable enough to be deployed in various real world applications, e.g. timetable or cinema information systems. The more conversational a dialogue system becomes, the less predictable are the users' utterances. Recognition and processing become increasingly difficult and unreliable. This is due to the fact that on virtually all levels in the natural language processing pipeline, ambiguities, underspecification and noise multiply greatly.

Research projects struggled to overcome the problems arising with more conversational dialogue systems, e.g. [Allen et al., 2000, Malaka and Porzel, 2000, Johnston et al., 2002, Wahlster, 2003, Boves, 2004]. Their goals are more intuitive and conversational natural language interfaces that can someday be used in real world applications. The work described herein is part of that larger undertaking as I view the handling of contextual - and therefore linguistically implicit - information & knowledge as one of major challenges for understanding conversational utterances in complex dialogue systems. For this we will outline the various ways of dealing with context proposed in the literature and how context-dependent processing has been implemented in systems that seek to understand natural language input.

As in different fields of linguistics, e.g. pragmatics, cognitive-, socio- and psycholinguistics, the relations between utterances and context are also of concern to computational approaches. These have to specify how to compute the relations between linguistic and contextual aspects. This is important for both natural language understanding as well as generation. In understanding the

Table 2.1: Employment of context in early dialog systems

Context	Usage
Domain Knowledge	lexicon building and syntactic categories
(static)	communicative function, e.g. speech acts
Discourse Knowledge	dialogue states and action planning
(dynamic)	reference resolution, e.g. anaphora

question is how to *decode* the context-dependent aspects of a linguistic expression. In generation one wants to encode contextual information into the linguistic expression.

2.4.1 From Past to Present: The Historical Context

With its beginnings in the 1960s the first NLU systems drew primarily on lexical and syntactic recourses and aimed at recognizing patters that had specific significances for the target applications. Semantics in those systems was constituted by the application-specific significances of certain words and phrases or domain-specific categories as elements of *semantic grammars*, e.g. the PLANES [Waltz, 1978] or LIFER/LADDER [Hendrix, 1977] systems. First considerations of contexts emerged with the first attempts to build more realistic NLU systems starting with SHRDLU [Winograd, 1972] and LUNAR [Woods, 1977]. In these systems syntactic and semantic rules were used to parse utterances into components and to compute the ensuing consequences for the system. Only SHRDLU performed some dialogue functions and some context-dependent analysis restricted to discourse context. Experimental systems hence have increased their capabilities and involved contextual analysis as shown in Table 2.1.

Visible in all these experimental systems that were limited to such an impoverished contextual analysis and precompilations, was their restrictedness in terms of understanding capabilities, rendering them unscalable and in the case of more conversational input undeployable. This evidently shows up in the fragility of systems that fail when confronted with imperfect or unanticipated input, usually that also includes perfectly unambiguous utterance that stray but a little from a scripted demo dialogue. As noted above human conversations are between partners that share a rich background of contextual knowledge (some more static & some more dynamic contexts) without which natural language utterances become ambiguous, vague and informationally incomplete.

An interpreter with little context awareness and interpretation will encounter problems and fail frequently; one which does not fail in unexpected or problematic situations is called *robust*. Several means have been used to increase robustness as listed in Table 2.2. These so-called low-level techniques [Bunt, 2000] have not solved the problem of enabling system to react felicitously in a dynamic context. These techniques fail to assume a pragmatics-based approach where fact that the user has an intention communicated via a message where the intend has to be reconstructed by recourse to the current context. Advances in contex-

Table 2.2: Means to increase robustness of early dialog systems

Object	Method
grammar	special rules and relaxations as well as automatic acquisition of semantic grammars
textual input	automatic spelling correction
lexica	on-line lexical acquisition

tual analysis have been implemented in a handful of systems that increased their capabilities and involved contextual analysis insofar as domain ontologies have been employed for lexicon building, syntactic categories and semantic parsing and discourse knowledge for modeling dialogue states, action planning as well as for resolving anaphora and ellipsis.

For example, the PHILQA system [Bronnenberger et al., 1997] featured context independent syntactic and semantic analyses as well as underspecified representations and context-dependent resolution with respect to the given domain representation. The SPICOS and TENDUM systems featured a resolution of structural ambiguity with underspecification, mass/count quantification with metavariables, and communicative functions determined by the user context [Bunt, 1984, Deemter et al., 1985]. Contextual underspecification was enabled by quasi-logical forms without semantic definition, which were instantiated unambiguously later by recourse to the semantic domain context, e.g. as implemented in the CLE system [Alshawi and Moore, 1992]. In much the same way the influential TRAINS and TRIPS systems used unscoped logical forms as well as speech acts with context represented as user/system beliefs [Allen et al., 1995, Ferguson and Allen, 1998]. While these systems put a main focus on spatial domains helping users to solve specific tasks and produced considerable progress through developing corpora and NLP components the main emphasis rested on the planning part of the system.

Other systems employed dialog acts and thematic structures to decontextualize underspecified semantic representations or logical forms, such as VERBMOBIL [Wahlster et al., 1993] and PLUS/DENK [Bunt, 1989]. Given the distinction between global (unchanging or hardly changing) context, i.e. domain/world knowledge and local (changing) context, about the situation, user beliefs, system intentions or discourse, contextual considerations have either looked at utterances as a whole [Searle, 1975, Allen and Perrault, 1986, Perrault, 1989, Ramsey, 2000] or focused on reference & anaphora resolution [Grosz et al., 1977, Webber, 1991, Byron, 2002, Poesio, 2002]. On a rather general level particular computational linguistic knowledge sources can be organized into context-variant and -invariant ones as shown in Table 2.3 [Porzel and Strube, 2002].

2.4.2 The Present: Multimodal Systems

Enormous contributions to the field of Computational Linguistics come from attempts that focus on aiding human-human communication. Research sys-

Table 2.3: Context-variant and -invariant levels of analysis

	Context-variant	Context-invariant
Speech Recognition	vocabulary language model	basic vocabulary
Syntax & Parsing	open class lexicon parsing	closed class lexicon grammar
Semantics	disambiguation domain knowledge	lexical semantics á la DRT common-sense knowledge
Pragmatics	intention recognition	dialogue acts

tems such as Verbmobil and the C-STAR translators or commercial systems such as the *Personal Translator* [Hahn and Amtrup, 1996, Cettolo et al., 1999, Bub and Schwinn, 1999] have created architectures, standards and principles which also feature discourse context-sensitive understanding of an utterance's meaning [Pinkal et al., 2000]. However, that is not always the same as understanding the underlying intention, when the system has to *answer* to this input. There are several academic and commercial tools available which include information extraction systems, information retrieval systems, knowledge acquisition systems, spell-checker, auto summarizer or dictation systems. Usually these tools are seen as components of NLP systems and not as systems on their own. Previous systems focusing on human-computer interaction are by and large either focused on sophisticating their natural language input (understanding) side or their output (production) side. An additional common characteristic of existing systems is that they are bound to single, specific domains and their employment of (a priori) defined scripts for dialog management. However, several end-to-end spoken dialog systems and multimodal research prototypes exist. Most notably, the TRAINS system and its successor TRIPS [Ferguson and Allen, 1998] constitutes such a spoken dialogue systems which attempts to help users to solve tasks. Though this attempt involved a considerable amount of work in developing corpora and NLP components, the main emphasis lies on the planning part of the system [Allen et al., 1996]. Also, both systems deal with tiny domains. The AT&T telephone-based system *May I help you?* [Gorin et al., 1997] is – like the majority of spoken dialogue systems coming out of AT&T – restricted to a single domain with not much more than a dozen conversational topics. The same is also true for EVAR [Gallwitz et al., 1998] and the Philips train timetable system [Aust et al., 1995].

Multimodal dialogue systems as the QuickSet system [Cohen et al., 1997], the Command Talk spoken language system [Stent et al., 1999], the EMBASSI system [Herfet et al., 2001] or the Match system [Johnston et al., 2002] are quite narrow in focus and coverage of speech input. The vocabulary of these systems covers only a few hundred entries and the domain knowledge contains only a few dozen concepts. These systems allow interaction only in a very controlled fashion. In general the follow identical architectural pipelines, which have been generalized in the EMBASSI framework as shown in Figure 2.3.

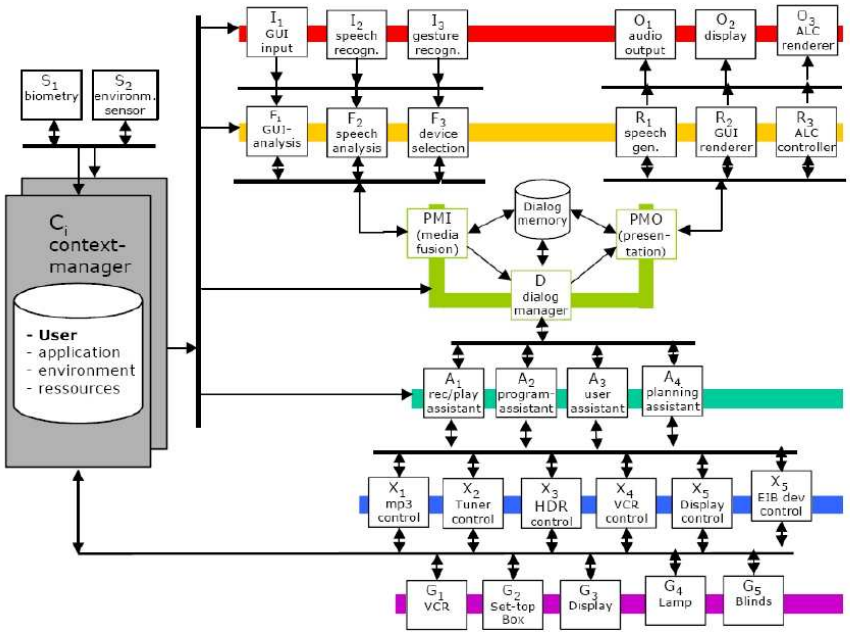

Figure 2.3: The EMBASSI multimodal architecture

This architecture consists of parallel input modalities (denoted by the letter I_n), which can be realized by automatic speech recognition systems, gesture recognition system or a graphical user interface. Their output is handed over to the respective modality-specific analyzers (denoted by the letter F_n). Correspondingly the modality-specific system responses are generated by a set of renderers (denoted by the letter R_n) and communicated via the modality-specific output mechanisms, such as graphical output or speech synthesis (denoted by the letter O_n). The task of multimodal fusion - unifying the input of the analyzers - and multimodal fission - distributing the output unto the renderers - is performed by the corresponding fusion and fission modules (denoted by PMI and PMO respectively). Ignoring the assistance and execution systems described in this architecture, the remaining part consists of a context manager, which obtains its input from biometric and sensoric input devices and stored information about the connected applications, devices in the environment and the user's preferences.

The SmartKom system [Wahlster et al., 2001] comprises a large set of input and output modalities which the most advanced current systems feature, together with an efficient fusion and fission pipeline. SmartKom features speech input with prosodic analysis, gesture input via infrared camera, recognition of facial expressions and their emotional states. On the output side, the system features a gesturing and speaking life-like character together with displayed gen-

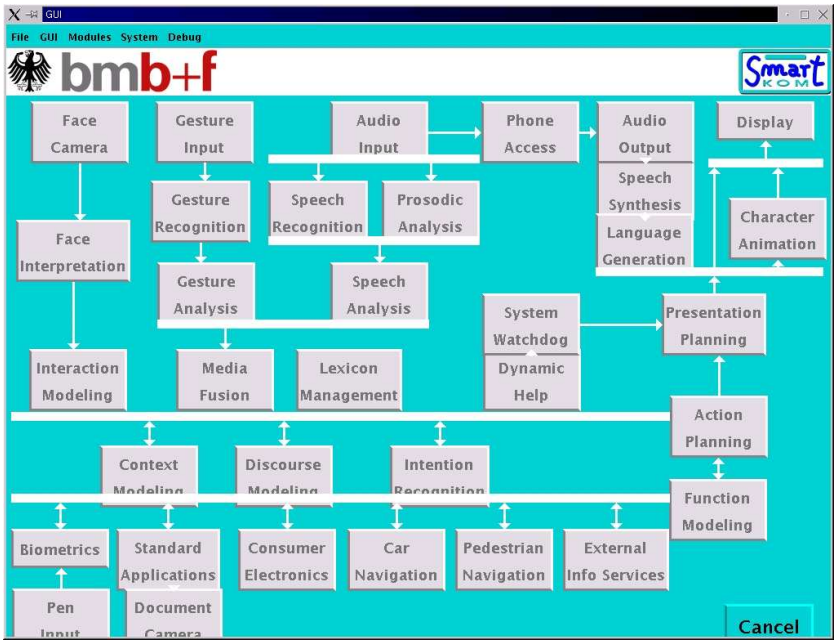

Figure 2.4: The SmartKom multimodal architecture

erated text and multimedia graphical output. It comprises nearly 50 modules running on a parallel virtual machine-based integration software called *Multiplatform*[3] and shown in Figure 2.4.

2.5 Methodological Background

In this section I will present methodological approaches for evaluating the performance dialog-, speech- and discourse understanding systems in the light of their pertinence for the evaluations performed in this work as well as their respective state of the art. Therefore, I will sketch out the most frequently used metrics for evaluating the performances of the relevant components and systems at hand in terms of their pertinence and applicability, focusing also on the specific contribution to this field that were brought about as a result of the measurements and metrics adopted in this work.

2.5.1 Performance in Dialogue Systems Evaluations

For evaluation of the overall performance of a dialogue system as a whole frameworks such as PARADISE [Walker et al., 2000] for unimodal and PROMISE

[3]The abbreviation stands for "MUltiple Language / Target Integration PLATform FOR Modules".

[Beringer et al., 2002] for multimodal systems have set a *de facto* standard. These frameworks differentiate between:

- dialogue efficiency metrics, i.e. elapsed time, system- and user turns
- dialogue quality metrics, mean recognition score and absolute number as well as percentages of timeouts, rejections, helps, cancels, and barge-ins,
- task success metrics, task completion (per survey)
- user satisfaction metrics (per survey)

These metrics are crucial for evaluating the aggregate performance of the individual components, they cannot, however, determine the amount of understanding *versus* misunderstanding or the system-specific *a priori* difficulty of the understanding task. Their importance, however, will remain undiminished, as ways of determining such global parameters are vital to determining the aggregate usefulness and felicity of a system as a whole. At the same time individual components and ensembles thereof - such as the performance of the uni- or multimodal input understanding system - need to be evaluated as well to determine bottlenecks and weak links in the discourse understanding processing chain.

2.5.2 Performance in Automatic Speech Recognition Evaluations

The commonly used word error rate (WER) can be calculated by aligning any two sets word sequences and adding the number of substitutions S, deletions D and insertions I. The WER is then given by the following formula where N is the total number of words in the test set.

$$WER = \frac{S + D + I}{N} \times 100 \tag{2.1}$$

Another measure of accuracy that is frequently used is the so called *Out Of Vocabulary* (OOV) measure, which represents the percentage of words that was not recognized despite their lexical coverage. WER and OOV are commonly intertwined together with the combined acoustic- and language-model confidence scores, which are constituted by the posterior probabilities of the hidden Markov chains and n-gram frequencies. Together these scores enable evaluators to measure the absolute performance of a given speech recognition system. In order to arrive at a measure that is relative to the given task-difficulty, this difficulty must also be calculated, which can be done by means of measuring the perplexity of the task see Section 2.6.

2.5.3 Performance in Understanding Evaluations

A measure for understanding rates - called *concept error rate* has been proposed for example by Chotimongcol and Rudnicky (2001) and is designed in analogy

to word error rates employed in automatic speech recognition that are combined with keyword spotting systems [Chotimongcol and Rudnicky, 2001]. They propose to differentiate whether the erroneous *concept* occurs in a *non-concept slot* that contains information that is captured in the grammar but not considered relevant for selecting a system action (e.g., politeness markers, such as *please*), in a *value-insensitive slot* whose identity, suffices to produce a system action (e.g., affirmatives such as *yes*), or in a *value-sensitive slot* for which both the occurrence and the value of the slot are important (e.g., a goal object, such as *Heidelberg*). An alternative proposal for concept error rates is embedded into the speech recognition and intention spotting system by Lumenvox[4], wherein two types of errors and two types of non-errors for concept *transcriptions* are proposed:

- A *match* when the application returned the correct concept and an *out of grammar match* when the application returned no concepts, or discarded the returned concepts because the user failed to say any concept covered by the grammar.

- A *grammar mismatch* when the application returned the incorrect concept, but the user said a concept covered by the grammar and an *out of grammar mismatch* when the application returned a concept, and chose that concept as a correct interpretation, but the user did not say a concept covered by the grammar.

Neither of these measures are suitable for our purposes as they are known to be feasible only for context-insensitive applications that do not include discourse models, implicit domain-specific information and other contextual knowledge as discussed in Porzel et al [Porzel et al., 2006a]. Therefore this measure has also been called *keyword recognition rate* for single utterance systems. Another crucial shortcoming noted [Porzel and Malaka, 2004b], is the lack of comparability, as these measures do not take the general difficulty of the understanding tasks into account. Again, this has been realized in the automatic speech recognition community and led to the so called *perplexity* measurements for a given speech recognition task. I will, therefore, sketch out the commonly employed perplexity measurements in Section 2.6.

The most detailed evaluation scheme for discourse comprehension, introduced by Higashinaka *et al* (2002), features the metrics displayed in Table 2.4 [Higashinaka et al., 2002]. Higashinaka *et al* (2003) combined these metrics by means of composing a weighted sum of the results of multiple linear regression and a support-vector regression approaches [Higashinaka et al., 2003]. This sum is. then, compared to human intuition judgments and metrics, comparable to PARADISE metrics [Walker et al., 2000], concerning task completion rates and -times. While this promising approach manages to combine factors related to speech recognition, interpretation and discourse modeling, there are some

[4] www.lomunevox.com/support/tunerhelp/Tuning/Concept_Transcription.htm

Table 2.4: Proposed measurements of discourse comprehension

1	slot accuracy
2	insertion error rate
3	deletion error rate
4	substitution error rate
5	slot error rate
6	update precision
7	update insertion error rate
8	update deletion error rate
9	update substitution error rate
10	speech understanding rate
11	slot accuracy for filled slots
12	deletion error rate for filled slots
13	substitution error rate for filled slots

shortcomings that stem from the fact that this schema was developed for single-domain systems that employ frame-based attribute value pairs for representing the user's intent.

Nevertheless, recent advances in discourse modeling, as described in Section 3.1.2 together with multi-domain systems enable approaches that are more flexible and more difficult to evaluate than the slot-filling measures described above, as they employ discourse pegs, dialogue games and overlay operations [Pfleger et al., 2002, Alexandersson and Becker, 2003] for handling more conversational input and cross-modal references . More importantly, no means of measuring the *a priori* discourse understanding difficulty is given, as I will discuss in Section 2.6.

2.5.4 Performance in Classification Evaluations

In the realm of semantic analyses the task of word sense disambiguation is usually regarded as the most difficult one. This means it can only be solved after all other problems involved in language understanding have been resolved as well. The hierarchical nature and interdependencies of the various tasks are mirrored in the results of the corresponding competitive evaluation tracts - e.g. the message understanding conference (MUC) or SENSEVAL competition. It becomes obvious that the ungraceful degradation of f-measure scores (shown in Table 2.5.4 is due to the fact that each higher-level task inherits the imprecisions and omissions of the previous ones, e.g. errors in the named entity recognition (NE) task cause recall and precision declines in the template element task (TE), which, in turn, thwart successful template relation task performance (TR) as well as the most difficult scenario template (ST) and co-reference task (CO). This decline can be seen in Table 2.5.4 that presents their corresponding f-measures - where precision and recall are weighted equally as given by the

Table 2.5: Evaluation results of the best systems of the 7th Message Understanding Conference

NE	CO	TE	TR	ST
f $\leq$ .94	f $\leq$ .62	f $\leq$ .87	f $\leq$ .76	f $\leq$ .51

Formula 2.2 below [Marsh and Perzanowski, 1999].

Despite several problems stemming from the prerequisite to craft costly gold standards, e.g. tree banks or annotated test corpora, precision and recall and their weighable combinations in the corresponding f-measures (such as given in Table 2.5.4), have become a *de facto* standard for measuring the performance of classification and retrieval tasks [Van Rijsbergen, 1979]. Precision p states the percentage of correctly tagged (or classified) entities of all tagged/classified entities, whereas recall r states the positive percentage of entities tagged/classified as compared to the normative amount, i.e. those that ought to have been tagged or classified. Together these are combinable to an overall f-measure score, defined as:

$$F = \frac{1}{\alpha\frac{1}{p} + (1-\alpha)\frac{1}{r}} \tag{2.2}$$

Herein α can be set to reflect the respective importance of p *versus* r, if $\alpha = 0.5$ then both are weighted equally. These measures are commonly employed for evaluating part-of-speech tagging, shallow parsing, reference resolution tasks and information retrieval tasks and sub-tasks.

An additional problem with this method is that most natural language understanding systems that perform deeper semantic analyses produce representations often based on individual grammar formalisms and mark-up languages for which no gold standards exist. For evaluating discourse understanding systems, however, such gold standards and annotated training corpora will continue to be needed.

2.6 Measuring Task Difficulties and Baselines

As the measurements, presented in Section 2.5.3, are not designed to reflect complexity of the tasks performed by the relevant components and systems at hand. I will, therefore, present the most frequently used metrics for estimating the difficulty inherent in such tasks as will be pertinent herein.

2.6.1 Measuring Perplexity in Automatic Speech Recognition

Perplexity is a measure of the probability weighted average number of words that may follow after a given word [Hirschman and Thompson, 1997]. In order

to calculate the perplexity B, the word entropy H needs to be given for the specific language of the system W. The perplexity is then defined within limits as:

$$0 < H = -\sum_{\forall 1<W<n} P(W) log_2 P(W) < log_2 n \tag{2.3}$$

$$B = 2^H$$

Improvements of specific speech recognition systems can then consequently be measured on a corpus with a given perplexity by measuring the corresponding error rates (WER and OOV) Together, this yields a performance measure for recognition quality that can be compared to other speech recognition performances on corpora with differing perplexity. The more common approach is to employ baseline measurements as a comparison for individual performances, e.g., where perplexity measures or other task-difficulty metrics are not at hand, as it is usually the case in classification tasks. I will, consequently, present pertinent baseline approaches in the following section.

2.6.2 Measuring Task-specific Baselines

Baselines for the performance of classification tasks are commonly defined based on chance performance, on an *a posteriori* computed majority class performance or against the performance of an established classification method. In other words, using the f-measure for performance discussed in Section 2.5.4, one can ask:

- what is the corresponding f-measure, if the evaluated component guesses randomly - for chance performance metrics,
- what is the corresponding f-measure if the evaluated component always chooses the most frequent solution - for majority class performance metrics,
- what is the corresponding f-measure of the established baseline classification method.

Much like kappa coefficiency measures for statistical inter-rater agreement, where observed agreement $P(a)$ is set in relation to what one would have expected $P(e)$ as shown in Formula 2.4 [Galton, 1892, Cohen, 1960, Carletta, 1996], existing employments of majority class baselines assume an equal set of identical potential mark-ups, i.e. attributes and their values, for all markables.

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)} \tag{2.4}$$

Therefore, they cannot be used in a straight forward manner for many tasks that involve disjunct sets of attributes and values in terms of the type and number of attributes and their values involved in the classification task. This, however,

Table 2.6: Summary of performance and difficulty measurements

Domain	Performance	Difficulty
automatic speech recognition	WER/OVV	Perplexity
natural language understanding	CER	none
MUC tasks (NE, TE, TR, ST, CO)	f-measure	baselines
unimodal dialogue system	PARADISE	none
multimodal dialogue system	PARADISE	none

is exactly what we find in natural language understanding tasks, such as in so-called *sense tagging* or *word sense disambiguation* tasks [Stevenson, 2003]. Additionally, baseline computed on other methods cannot serve as a means for measuring scalability, because of the circularity involved: as one would need a way of measuring the baseline method's scalability factor in the first place. Table 2.6.2 provides an overview of the existing ways of measuring performance and task difficulty in automatic speech recognition and understanding.

Current evaluation frameworks for uni- or multimodal dialogue systems [Walker et al., 2000, Beringer et al., 2002] that allow for spoken language input do not include metrices for measuring the accuracy of the involved intention recognition systems, simply because such information is hard to extract automatically from the logs of system runs [Litman et al., 1999b]. Furthermore, no general computational method or framework for measuring the difficulty of natural language understanding tasks have been proposed so far. We are, therefore, faced with a lack of methods for measuring the difficulties of the individual tasks involved in the language understanding process. Such generally applicable methods, however, are needed for measuring the scalability of natural language understanding systems and components.

2.7 Point of Departure

Utterances in dialogues, whether in human-human interaction or human-computer interaction, occur in a specific situation that is composed of different types of contexts. In the following a categorization of the types of context relevant

to spoken dialogue systems - and human computer interaction in general - is given together with their respective scope (content) and modularization in HCI systems. This work will, as observable below, depart from the common distinction between linguistic and extra-linguistic contexts, whereby all extra-linguistic contexts are also often lumped together as the *situational context* [Connolly, 2001]. The categorization proposed herein subsumes the linguistic context under the heading of *dialogical context*. Dialogical context encompasses the dialogical counterparts of both co-text and inter-text as well as non-linguistic input from other modalities, e.g. interacting with traditional interfaces (WIMP) or gesture and the like. The categorization employed herein also differentiates extra-linguistic (or situational) context into interlocutionary-, domain- and situational context as shown below.

2.7.1 Context Types

As shown in table 2.7 dialogical context in our model corresponds to what has been termed *linguistic context* in the domain of natural language analysis and encompasses information from the discourse history, i.e. prior utterances by the interlocutors. As pointed out in section 2 discourse context is essential for a variety of tasks that one finds under the headings of *reference resolution*, *anaphora resolution* or *semantic disambiguation*. The following section presents an elaboration of these problems in the light of the essential contributions of context.

Table 2.7: Contexts, content and knowledge sources

types of context	content	knowledge store
domain context	world/conceptual knowledge	domain model
dialogical context	what has been done by whom	dialogue model
interlocutionary context	properties of the interlocutors	user model
situational context	time, place, etc	situation model

2.7.2 The Tasks (Revisited)

In order to employ a consistent terminology in the subsequent discussions and experiments on finding appropriate meanings for given linguistics forms, I will adopt - wherever possible- the basic notations and insights that originated in the so called *construction grammar* framework [Lakoff, 1987, Langacker, 1987, Fillmore, 1988, Talmy, 1988]. As also noted in Section 4 work on context and real language use in formal linguistics was based the earlier insights in functional and usage-based models of language and was mainly restricted to the field of Cognitive Linguists.

The ensuing grammatical framework and vocabulary in formal construction grammar [Goldberg, 1995, Kay and Fillmore, 1999, Feldman, 2006], has been explicitly devised to handle actually occurring natural language phenomena,

which notoriously contains non-literal, elliptic, context-dependent, metaphorical or underspecified linguistic expressions. As shown in this chapter, these phenomena still present a real challenge for current natural language understanding systems. Furthermore, I agree with the central principle of construction grammar which states that grammatical phenomena also contribute to the meaning of a sentence which is the reason why syntax cannot be defined independently of semantics of a grammar.

Constructions are the basic building blocks, posited by the construction grammar framework, and are defined as follows [Goldberg, 1995]: A construction is a form-meaning pair $< F_i, S_i >$ if some aspect of F_i or some aspect of S_i is not strictly predictable from the component parts of that construction or from other previously established constructions. (Ibid:4). Using this framework, the aforementioned task of resolving referring expressions and ambiguities can be stated as follows: Given a form F_i, which can be a referring expression such as *the bank* or an anaphora such *it*, what is the corresponding meaning $S(F)$ in the given context. It is important to keep in mind that a form in construction grammar can be constituted on all linguistic levels, i.e. we find phonological forms, morpho-syntactic forms, lexical forms and clausal forms. That means, one can describe lexical constructions as *the* and *bank* individually, look at a composite construction such as the referring expression *the bank* or even a whole utterance such as *Where is the bank* and how they can *resolved*, meaning which specific meaning is to be assigned to it given the context at hand.

Computationally, this entails - as I will show in Sections 3.2 through 3.5 as well as in Section 4.2 - dealing with our target challenges in automatic language understanding for resolving a contextually adequate formal specification of the semantics from the given ensemble of forms. The target data structure will, consequently, be referred to as a *semantic specification* [Chang et al., 2002]. The corresponding tasks in natural language generation are selecting (constructing) in a context-dependent manner - a semantic specification out of myriads of alternatives and the ensuing construction-based formulation thereof.

Numerous works have sought to label various relations between forms and meanings. I have already exemplified the difference between homonymy and polysemy, but additional phenomena have received great attention, e.g. metonymy [Hobbs, 1991, Markert, 1999], metaphor [Lakoff and Johnson, 1980], coercion [Michaelis, 2001], type shifting [Faucounnier and Turner, 1998] and mental spaces [Fauconnier, 1985]. While this work will not discuss these phenomenon in greater detail, it is important to note that the fundamental assumption underlying such analysis is that individual forms feature some kind of literal meaning and that they can assume non-literal meaning by means of metonymical or metaphorical usage, coercion and the like. While the work presented herein departs from this assumption, our view is not irreconcilable with it.

Various terms have been proposed in the literature, e.g., *intricacy* or entrenchment [Faucounnier and Turner, 1998, Langacker, 2000], that express that these phenomena can be measured on a scale. Which means, in the words of Faucounnier and Turner, that meaning can be assigned to forms with increasing or decreasing intricacy. A *literal* usage would require little to no intricacy and oth-

ers, such as so-called *blends* require more (ibid). For Langacker this intricacy can be boiled down to statistical measure of *entrenchment*, the more frequently used a specific contextually evoked form - meaning pairing becomes, the more central is the correspondingly entrenched meaning of that form [Langacker, 2000].

Before returning to these questions below, let me point out, once more, that the general task of determining a particular meaning representation - that has to be constructed with more or less intricacy from the forms at hand - will be the central empirical domain to be employed in the approach to contextual computing presented herein. If linguistic forms were to be unambiguous and always fully specific, then no additional meaning construction would be necessary due to the given one-to-one mapping between a specific form and its meaning. Since that is, obviously, not the case additional information and knowledge is needed to construe the intended meaning of a given form.

In the computational sense this entails that meaning resolution can be seen as determining the most plausible meaning from the set of the possible meanings that can be constructed out of the given form.[5] In line with the central claim of this work, we, therefore, find that on all computational levels of natural language processing where underspecification, ambiguity and noise arises, one needs additional information and knowledge for finding the most plausibly constructed meaning, thereby resolving the form-meaning mappings out of many other potential ones. As I will argue below contextual information and pragmatic knowledge constitutes this additional modality by means of which meaning resolution becomes possible, or - in other words - the other potential form-meaning mappings are inhibited from being activated as the most plausible one did.

What would, therefore, be needed is a context-dependent scoring that identifies the most plausible item out of a set of possible alternatives. In the following sections I will introduce, examine and evaluate how such an approach to contextual computing can be employed to increase performance, robustness and scalability of natural language processing systems in the areas of:

- Automatic Speech Recognition
- Semantic Interpretation
- Pragmatic Interpretation

Before I present the data, experiments and results of applying the approach to contextual computing pursued herein to these areas of natural language processing, I will discuss the subsequent modeling of contextual knowledge stores employed by the contextual computing approach to be discussed hereafter. Since domain knowledge is nowadays commonly modeled using formal ontologies, they will be introduced first generally and then specifically, in terms of the concrete ontologies and modeling principles employed to represent domain knowledge in our experiments.

[5] This, in turn, is quite congruent to other approaches in contextual computing where the computational notion of *correctness* is - by necessity - replaced with the notion of *plausibility*.

Chapter 3

Domain and Discourse

3.1 Modeling Domain and Discourse Knowledge

The early ways in which knowledge has been represented semantically in spoken dialogue systems or multi-modal dialogue systems show that individual representations with different semantics and heterogeneously structured content can be found in various formats within single natural language processing systems and applications. For example, a typical NLP system, such as TRAINS [Allen et al., 1996], employs different knowledge representations for parsing, action planning and generation, despite the fact that what is being represented is common to all those representations, e.g., the parser representation for *going from A to B* has no similarity to the action planner's representation thereof [Ferguson et al., 1996]. Also central concepts, for example *city*, are represented semantically in multiple ways throughout the system.

The origin for this early state of affairs is that the respective knowledge stores were hand-crafted individually for each task. Sometimes they are compiled into code and cease to be externally available. Where an explicit knowledge representation is used, one finds a multitude of formats and inference engines, which often cause both performance and tractability problems. In this section I will, therefore, introduce representational formats for formal ontologies followed a description of the knowledge representation i.e. the formal ontology, used to serve as a representation of the domain context within complete multi-modal dialogue system. Therefore, I will describe the underlying modeling principles and the benefits of such a rigorously crafted knowledge store.

3.1.1 Modeling Domains

Recently developed multi-modal dialogue systems equipped with the ability to understand and process natural language utterances from one or more domains [Wahlster et al., 2001, Johnston et al., 2002, Reithinger et al., 2005] employ formal ontologies as defined notoriously by Gruber (1993) as a **conceptual specification** of a domain of interest [Gruber, 1993] which is:

- machine-readable - formal;
- semantics are based on logic - explicit.

Formal ontologies have already been employed to represent domain-specific semantic specifications in natural language processing systems [Chang et al., 2002, Porzel et al., 2006b]. Before showing examples of semantic domain specifications, I will present the formal representations that have been emerging, e.g. from the Semantic Web Project [Berners-Lee et al., 2001], that employ such formal conceptualizations to add semantic information to textual and other data available on the Internet [Baader et al., 2003]. Efforts originating in web standardization consortia and projects, such as the World Wide Web Consortium [W3C-OEP, 2005] and the aforementioned Semantic Web project, brought about a series of knowledge modeling standards based on the - ever more adequately named - Extensible Markup Language (XML) [XML, 2001]. Most notably, for this work the Resource Description Framework (RDF) and the corresponding formal means to define vocabularies and grammars for RDF instances such as Resource Description Framework Schemata (RDFS) [RDF, 2001, RDFS, 2001]. Building upon this framework specific proposals included the DARPA Agent Mark-up Language (DAML) [Klein et al., 2000], Ontology Interchange Language (OIL) [Fensel et al., 2001] and their mix (DAML+OIL) [Gil and Ratnakar, 2002] resulting in the Ontology Web Language standards - OWL-Lite, OWL-DL, OWL-Full and OWL2 - that have subsequently been employed for crafting knowledge representations ranging from so-called *lightweight* ontologies, e.g., mere taxonomic models, to fully axiomatized representations of foundational, domain-independent and domain-specific ground and descriptive knowledge [Gangemi et al., 2002].

It has been shown that foundational and domain-specific knowledge - based on RDFS grammars and vocabularies - can be employed for representing knowledge in multimodal dialog systems [Gurevych et al., 2006, Oberle et al., 2007]. The expressive capabilities of the grammars that have been proposed vary [Fensel et al., 2001], this is due to typological and grammatical variations of ontology *dialects*. The ontologies implemented in the corresponding SmartKom and SmartWeb systems are described in greater detail in Sections 4.3 and 3.1.3.[1]

Formal ontological models of domain knowledge are also based on a formal semantics, which - corresponding to the logical calculus employed - enables specific reasoning engines, such as FACT, RACER or CEL [Horrocks, 1998, Haarslev and Möller, 2003, Baader et al., 2006]. They provide inferencing capabilities [Guarino and Welty, 2002], such as class consistency or subsumption checking. Also graphical ontology engineering front-ends and visualization tools are available for editing, maintaining, and visualizing the corresponding ontologies.

[1] For the time being it should suffice to note that using such an XML-based semantic mark-up language where instances can be defined in the syntax of the Resource Description Framework brings about several advantages, e.g., that the ontology instances are understandable for all RDF-based applications and other grammars that are constructed using RDFS.

Computationally speaking, the foundation of RDF is a model for representing named properties and property values. The RDF model draws on well-established principles from various data representation communities. RDF properties may be thought of as attributes of resources and in this sense correspond to traditional attribute-value pairs. RDF properties also represent relationships between resources and an RDF model can therefore resemble an entity-relationship diagram. In object-oriented design terminology, resources correspond to objects and properties correspond to instance variables.

The RDF data model is a syntax-neutral way of representing RDF expressions. The data model representation is used to evaluate equivalence in meaning. Two RDF expressions are equivalent if and only if their data model representations are the same. This definition of equivalence permits some syntactic variation in expression without altering the meaning. The basic data model consists of three object types:

- Resources: All things being described by RDF expressions are called resources. A resource may be an entire Web page; such as the HTML document "http://www.w3.org/Overview.html" for example. A resource may be a part of a Web page; e.g. a specific HTML or XML element within the document source. A resource may also be a whole collection of pages; e.g. an entire Web site. A resource may also be an object that is not directly accessible via the Web; e.g. a printed book. Resources are always named by URIs plus optional anchor ids (see [URI]). Anything can have a URI; the extensibility of URIs allows the introduction of identifiers for any entity imaginable.

- Properties: A property is a specific aspect, characteristic, attribute, or relation used to describe a resource. Each property has a specific meaning, defines its permitted values, the types of resources it can describe, and its relationship with other properties. This document does not address how the characteristics of properties are expressed; for such information, refer to the RDF Schema specification).

- Statements: A specific resource together with a named property plus the value of that property for that resource is an RDF statement. These three individual parts of a statement are called, respectively, the subject, the predicate, and the object. The object of a statement (i.e., the property value) can be another resource or it can be a literal; i.e., a resource (specified by a URI) or a simple string or other primitive data type defined by XML. In RDF terms, a literal may have content that is XML markup but is not further evaluated by the RDF processor. There are some syntactic restrictions on how markup in literals may be expressed;

Resources are identified by a resource identifier. A resource identifier is a URI plus an optional anchor id. For the purposes of this section, properties will be referred to by a simple name.

The semantics of ontological formalism is by and large based on description logic extended with concrete data types. The languages, therefore, employ a combination of frame- and description logic. It provides most of the modeling primitives commonly used in the frame-based knowledge representation systems. Frames are used to represent concepts. These frames consist of a collection of classes along with a list of slots and attributes. Under the term *class* or *class expression* a class name, or an enumeration, or a property-restriction, or a Boolean combination of class expressions is to be understood. Slots are interpreted as a collection of properties. They are divided into those that relate classes to other classes (so called *object properties*) and those that relate classes to data type values (so called *datatype properties*). Slots can be filled by: class names, names of the atomic elements, collection of the above (conjunctive sets - *and*, disjunctive sets - *or*, or negation - *not*), concrete data types (*integers* and *strings*).

Then, domain and range restrictions of the slots can be defined. Domain restriction asserts that the property only applies to the instances of particular class expressions. Range restriction specifies that a property only assumes values that are instances of the respective class expressions. Slot fillers can have several types of further constraints, also called *facets*. These include *value-type* restrictions (all fillers must be of a particular class), *has-value* restrictions (there must be at least one filler of a particular class). The *value-type* restriction corresponds to the universal quantifier of the predicate logic. The *has-value* restriction is analogous to the existential quantifier. Another constraint on the slot fillers is *cardinality*, which limits the number of possible fillers of the given class. Atomic elements or individuals can also be associated with a class definition via slot constraints.

3.1.2 Modeling Discourse

Most dialogue systems, and certainly multi-domain ones [Johnston et al., 2002, Reithinger et al., 2003], employ a discourse model that contains representations of objects mentioned throughout the discourse along with a record of what has actually been said by whom, which I will call the *discourse protocol*. Of course, linguistic investigations on discourse phenomenon, for example on anaphoric expressions that refer to objects in the discourse model that have been mentioned before are both numerous and diverse [Searle, 1975, Grosz et al., 1977, Lambrecht, 1994, Ramsey, 2000, Kehler, 2002] Generally, these co-referring expressions - of which there is a record in the discourse protocol - are called *antecedents*. I have shown the results of corpus-based analyses of the form sides of referring expression, such as *The first man on the moon*, in Section 2.3 [Poesio and Vieira, 1998, Byron, 2002, Poesio, 2002].

Using the formal ontological terms I presented in Section 3.1.1, representations of the referent, i.e. the entity to which the expression refers, in the discourse model can be an instances of some type, as in the case of named entities such as the individual person named Neil Armstrong, who is the referent of the referring expression given above. Additionally, hypernyms of named en-

tities[2] refer to terms in the so-called *T-Box* or the so-called *vocabulary* of the ontology, as in the case of the two referents of the two common noun constructions found in the expression: *Cats sit on mats*. Lastly, there a cases such as discourse deictic expressions and abstract anaphora where reference is made to parts of the discourse protocol as in *Could you repeat that* in the former and to referents that have been introduced via entire phrasal constructions in the later case. In both cases subsequent anaphoric shortening has also been observed [Webber, 1991, Byron, 2002].

Traditionally, formal discourse models have been discussed for uni-modal settings [Webber, 1979] while more recently formal models have been proposed for multimodal discourse modeling [LuperFoy, 1992, Alexandersson et al., 1995, Pfleger et al., 2003]. In her proposal LuperFoy suggest a three tiered model. Therein a surface layer, called the *linguistic layer* (ibid) the input is represented in terms of *linguistic objects*. This syntactically structured discourse protocol features a build-in decay as utterances lose relevance as time goes on. The inner layer that represents a model of the world, e.g. an ontology, is referred to more traditionally by LuperFox as the *knowledge base*. Between these two layers the resides the *discourse layer* in her model, which contains *amodal* representations of the referents of the ongoing discourse called *discourse pegs*. In case a new linguistic object comes into the surface buffer it is checked, if the instance or concept can be linked to an existing peg, as in the case of anaphoric relationships; if it can not be linked to an existing peg, it is assumed to refer to a discourse new referent. Given sufficient referential information pegs are linked to concrete object in the world model, knowledge base or ontology.

A shortcoming of the LuperFoy Model is that all non-linguistic modalities, such as gestures, must actually represented as linguistic objects in the linguistic layer. This, however, was remedied by a corresponding enhancement that included other modality-specific objects to be represented as such, e.g., visual objects as item presented to the user or gestural objects as items pointed out to the user or by the user [Pfleger et al., 2003]. This enhanced multimodal discourse model, displayed in Figure 3.1 was also employed in the SmartKom system thereby connecting the represented discourse objects (corresponding to LuperFoy's pegs) to the ontology, that I will describe below in Section 3.1.3. A tight coupling of representations employed in the domain model and the multimodal discourse model not only enabled the experiments presented in Section 3.3 [Porzel et al., 2003a], but also additional algorithms, such as *default unification* or *overlay* algorithms [Carpenter, 1992, Alexandersson and Becker, 2001, Löckelt et al., 2002], that can *hand down* logically fitting defaults or prior discourse objects. Apart from the discourse layer, the coupling of the domain model representations to those employed by the modality-specific layers also enabled multimodal fusion and fission algorithms in the SmartKom system [Porzel et al., 2003b]. Due to its central importance also for experiments presented in Sections 3.1 through 3.4, I will describe this specific inner layer, i.e.

[2]Proper names, as the lexical leafs of the hierarchy are by definition solely hyponyms of some general term, while all general terms are both the hypernyms of their hyponym as well as the hyponym of their more general term.

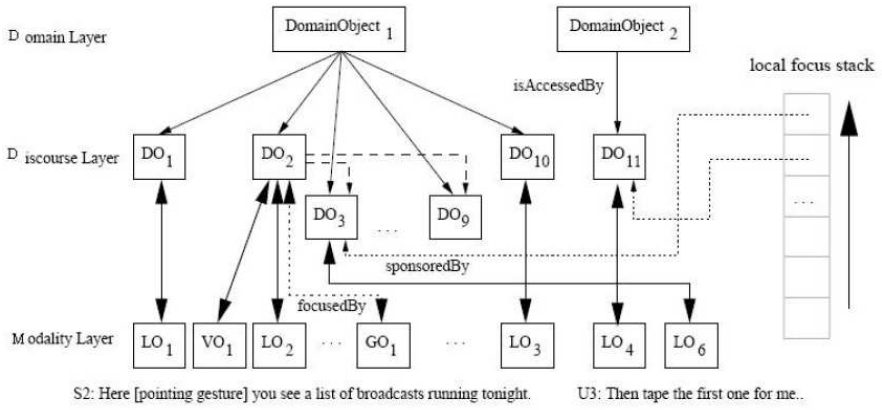

Figure 3.1: The SmartKom Discourse Model

the domain model, and its modeling principles in greater detail below.

3.1.3 Semantics in SmartKom

As mentioned above multimodal systems, such as SmartKom, can benefit from basing their semantic representations on ontological domain models. The specific model described herein, serves as a common knowledge representation for various modules throughout the system [Gurevych et al., 2003b]. It represents a general conceptualization of the world using a so-called *foundational* ontology [Russell and Norvig, 1995, Gangemi et al., 2002] as well as of particular domains (domain-specific ontologies). This way, the ontology represents language-independent world-knowledge. The language-specific knowledge can be stored elsewhere, e.g. in a lexical resource that links lexical forms to a semantic representation, e.g., defined in terms of ontology concepts or language-specific knowledge can be modeled and added as an ontology proper [Buitelaar et al., 2006].

The ontology to be employed in the subsequent experiments was initially designed as a general purpose component for knowledge-based natural language processing. It bases on a foundational ontology developed following the specific procedures proposed by Russel and Norvig (1995) and originally covered the tourism domain encoding knowledge about sights, persons and buildings [Russell and Norvig, 1995]. Then, the existing ontology was adopted in the SmartKom project and modified to cover a number of new domains, e.g. new media and TV program guides. The top-level (i.e. foundational and/or domain-independent) part of ontology was re-used with some slight extensions. Further developments were motivated by the need of a *process hierarchy*. This hierarchy models processes which are domain-independent in the sense that they can be

relevant for many domains, e.g. *InformationSearchProcess*

The ontology employed herein has about 730 concepts and 200 non-taxonomic relations. It includes a generic top-level ontology whose purpose is to provide a basic structure of the world, i.e. abstract classes to divide the universe in distinct parts as resulting from the ontological analysis. This top-level was developed following the procedure outlined in [Russell and Norvig, 1995]. The acquisition of the ontology went in two directions: top-down to create a top level of the ontology and bottom-up to satisfy the need of mapping lexical items to concepts. The purpose of the top-level ontology is to provide a basic structure of the world, i.e. abstract classes to divide the universe in distinct parts as resulting from the ontological analysis [Guarino and Poli, 1995]. The domain concepts emerged through a comprehensive analysis of collected corpus of multimodal human utterances [Schiel et al., 2002, Schiel and Türk, 2006].

3.1.4 Modeling Ground Knowledge

Following the top-level distinctions mentioned above, a collection of concepts that have *primary* ontological status [Guarino and Welty, 2000] were defined. The guiding principle was to differentiate between the primary ontological entities and the roles taken by them in particular situations, events, or processes. For example, a building can be a hospital, a railway station, a school, etc. But while taking all these roles, it doesn't cease to be a building. Another example is a person who can take the role of a school teacher, a mother, etc., but it still remains a person for.

Here the question arises, how deep the differentiation should go. Consider the example of a person: we give a concept *Person* a primary ontological status, but what about the concepts *Man* and *Woman*? Should they be given the same status? Our answer is positive and is based, on one hand, on the assumption that sex is the primary property that defines a person as a man or a woman, on the other hand, a functional approach shows that relations of these two classes to other classes and their other attributes can be determined by this property. In this way, the basic top-level ontological categorization in our system divides all concepts into two classes *Type* and *Role* (see Figure 3.2). As the class *Type* includes concepts with primary ontological status independent of the particular application, every system using the ontology for its specific purposes deals with the class *Role*. We will return to this question in later sections that discuss dedicated formal models of pragmatic knowledge

The Taxonomy and Vocabulary of the Ontology

Role is the most general class in the ontology representing actual roles that any entity or process can perform in a specific domain. It is divided into *Event* and *AbstractEvent*. In the view of the ontology employed herein, `Role` therefore represents a role that any entity or process can perform.

Along with events, e.g. processes or physical objects that exist in space or in time, our model includes abstract objects, e.g., numbers, abstract properties,

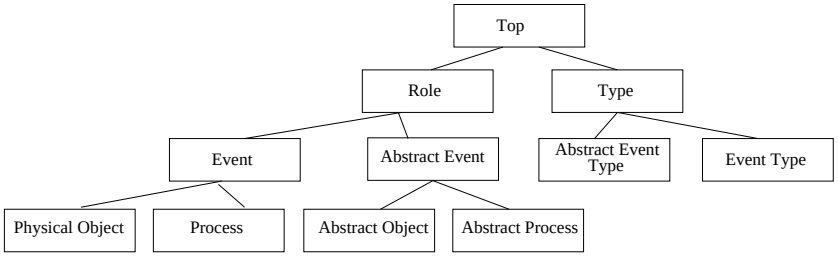

Figure 3.2: Top-level part of the ontology

such as spatial relations and abstract events relevant to real processes, such as *Start*, *Finish*, *Interrupt*, etc. These are modeled separately thereby allowing these modeled *patterns* in the description of the processes throughout the ontology. The class *AbstractEvent* further differentiates abstract object from abstract process in correspondingly named classes.

In the non-abstract domain *Event*'s themselves are further classified in *PhysicalObject* and *Process*. In contrast to abstract objects, they have a location in space and time. The class *PhysicalObject* describes any kind of objects we come in contact with - living as well as non-living. These objects correspond to roles in different domains, such as *Sight*, *Goal* and *Landmark* in the tourism domain or *Film* and *Movie* in the TV and cinema domain, etc., and can be associated with via semantic relations to the processes via slot constraint definitions. Next to events such as physical objects there processes defined in the ontology.

The modeling of *Process* as a kind of event that is continuous and homogeneous in nature, follows the frame semantic analysis used for generating the FrameNet data [Baker et al., 1998]. Based on the analysis of additional data collected by Schiel et al (2004) in the domains under inspection herein, we developed the classification of processes given in Figure 3.3.

- `General Process`, a set of the most general processes such as duplication, imitation or repetition processes;
- `Mental Process`, a set of processes such as cognitive, emotional or perceptual processes;
- `Physical Process`, a set of processes such as motion, transaction or controlling processes;
- `Social Process`, a set of processes such as communication or instruction processes.

While the three last classes can be understood intuitively, the first one needs further explanation. It consists of several subclasses, such as *AbstractDuplicationProcess*, *AbstractRepetitionProcess*, *AbstractImitationProcess*, etc. These are abstract processes that are independent from the real processes and can take

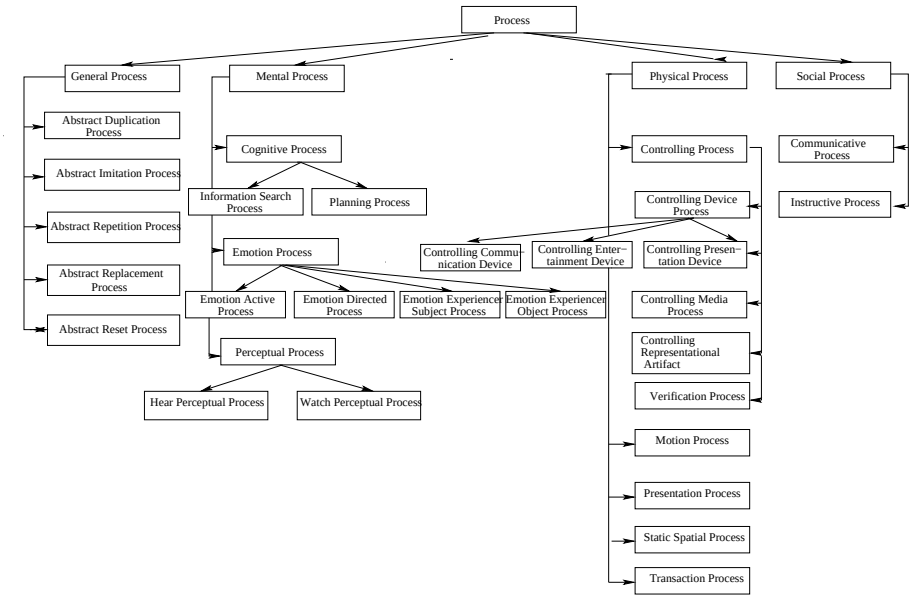

Figure 3.3: Upper part of the process hierarchy

place at the same time with the main process. For example, the semantic structure of an utterance "I used to come here" can be represented as a combination of the *Motion Directed Transliterated Process* and *Abstract Repetition Process*.

The *MentalProcess* subtree includes *CognitiveProcess*, *EmotionProcess* and *PerceptualProcess*. Under *CognitiveProcess* we understand a group of processes that aim at acquiring information or making plans about the future. The further division of *EmotionProcess* into the following subclasses - *EmotionExperiencerObjectProcess* and *EmotionExperiencerSubjectProcess* - is due to the fact that an emotion can be either provoked by an object (e.g. The cry scared me) or can be experienced by an agent towards some object (e.g. I want to go home).

The *PhysicalProcess* has the following subclasses: the semantics of *ControllingProcess* presupposes the controlling of a number of Artifacts, e.g. devices, *MotionProcess* models different types of agent's movement regarding some object or point in space, *PresentationProcess* describes a process of displaying some information by an agent, e.g. a TV program by *Smartakus*, an artificial character embedding the SmartKom system, *StaticSpatialProcess* consists in the agent's dwelling in some point in space, *TransactionProcess* presupposes an exchange of entities or services among different participants of the process.

Another subclass of the *Process* - *SocialProcess* includes *CommunicativeProcess*, which consists in communicating by the agent a message to the addressee by different means, and *InstructiveProcess* which describes an interaction between an agent and a trainee.

Let us consider the definition of the `Information Search Process` in the

ontology. It is modeled as a subclass of the `Cognitive Process`, which is a subclass of the `Mental Process` and inherits the following slot constraints:

- **begin time**, a time expression indicating the starting time point;
- **end time**, a time expression indicating the time point when the process is complete;
- **state**, one of the abstract process states, e.g. start, continue, interrupt, etc.;
- **cognizer**, filled with a class *Person* including its subclasses.

Information Search Process features one additional slot constraint, **piece-of-information**. The possible slot-fillers are a range of domain objects, e.g. *Sight*, *Performance*, or whole sets of those, e.g. as for *TV Program*, but also processes, e.g. *Controlling TV Device Process*. This way, an utterance such as shown in Example 11 can also be mapped onto *Information Search Process*.

(11) I would like information about the castle

This process has an agent of type *User* and a piece of information of type *Sight*. *Sight* has a name of type *Castle*. Analogously, the utterance shown in Example 12.

(12) How can I control the TV

can be mapped onto *Information Search Process*, which has an agent of type *User* and has a piece of information of type *Controlling TV Device Process*.

The class *Physical Object* describes any kind of objects we come in contact with - living as well as non-living - having a location in space and time in contrast to abstract objects. These objects refer to different domains, such as *Sight* and *Route* in the tourism domain, *AV Medium* and *Actor* in the TV and cinema domain, etc., and can be associated with certain relations in the processes via slot constraint definitions.

The Hierarchy of Non-taxonomic Relations

The structure of the non-taxonomic relations also reflects the general intention to keep abstract and concrete elements apart. A set of most general properties has been defined with regard to the role an object can play in a process: as an agent, theme, experiencer, instrument, means, location, source, target or path. These general roles applied to concrete processes may also have subslots: thus an agent in a process of buying (*TransactionProcess*) is a buyer, the one in the process of cognition is a cognizer. This way, non-taxonomic relations can also build a hierarchy. The property *has-theme* in the process of information search is a required *piece-of-information*, in presentation process it is a *presentable-object*, i.e. the item that is to be presented, etc.

Consider the class *Process*. It has the following relations: *has-beginning*, a typed time expression which here (plays the role) of indicating a starting point,

has-end, the same indicating the time point when the process is complete, *has-state*, one of the abstract process states. These relations describe properties that are common to all processes, and as such they are inherited by all subclasses of the *Process* class. The choice was made not to include *has-agent* into the number of properties of the *Process* class to account for the fact that not all processes have one.

An *EmotionExperiencerSubjectProcess*, for example, inherits the slots of the *Process* class, among them the slot *has-theme* that can be filled with any process or object (the basic idea is that any physical entity or the performance of any process can become an object of someone's emotion). It also has several additional properties such as *experiencer* to denote the one who undergoes the process, and *preference* to define the attitude an experiencer has to the object of its emotion. Any instance of the class *Person* (and its subclasses) can fill the slot *has-experiencer*, the filler of the slot *has preference* is any instance of the class *Attitude* (or any of its subclasses). *Attitude* is a subclass of the class *Abstract Object*, i.e., positive or negative attitude is modeled as an abstract object and then brought into connection with a certain process, e.g., the *Emotion Process*.

Another example demonstrating how slot structures can be shared between some super- and subclasses can be seen in the subclass *AvEntertainment*, that inherits from its superclass *Entertainment* the following slots: *has-duration*, *has-endtime*, and *has-begintime*, filled by the instances *TimeDuration* and *TimeExpression* respectively. The class *AvEntertainment* features two additional slots: *has-language*, its filler must be an individual instance of the class *Language*, e.g., English, and *has-medium*, its filler is of class *Medium*. The class *AvEntertainment* has further subclasses - *Broadcast* representing an individual entry in a TV program, and *Performance* modeling an entry in a cinema program. Both of them inherit the slots of the superclasses *Entertainment* and *AvEntertainment*, while also featuring their own additional slots, e.g., *has-channel* and *has-showview* for the *Broadcast*, *cinema* and *has-seat* for the *Performance*.

Given this elaborate structure of the non-taxonomic relations, it is important to note its significance for the contextual computing experiments - that I will describe in Sections 3.2 through 3.6 - as well as in the light of the correspondence between these non-taxonomic relations and so-called *role* relations in lexicographic frame semantic analysis [Fillmore and Baker, 2000, Rosario et al., 2002] as well as in all forms of role binding in constructional analysis, e.g. involving multiple specific role bindings afforded by argument structure constructions [Goldberg, 1995].

3.1.5 Roadmap

As I have discussed in Section 2.7.2, one of the fundamental task in natural language processing for dialogue systems is to pick the most plausible item out of a set of possible alternatives. As also noted earlier in Section 1.1, these alternatives multiply in conversational spoken dialog system due to noise, ambiguities, and underspecification, which is why, in the following, I will examine

Table 3.1: An overview of the areas and problems addressed

problem	application
noise in automatic speech recognition	hypothesis verification
ambiguities in semantic interpretation	word sense disambiguation
underspecification in pragmatic interpretation	pragmatic ambiguities

the contribution of contextual information & knowledge by looking at each of these problems in different areas of natural language processing, as shown in Table 3.1.

3.2 Using Domain Context for Noisy Input

Given such ontological models of domain knowledge, as described in Section 3.1 above, I will begin this examination of the contribution of using such domain knowledge for contextual computing by presenting a set of experiments performed to examine the effects of including domain context into the speech processing pipeline. More specifically, I will describe the data, the annotation thereof and the results concerning speech recognition ambiguities in the form of n-best on each of these tasks given in Table 3.2. I will, therefore, shortly introduce the models currently employed in speech recognition systems for the task of hypothesis verification.

3.2.1 The Task: Domain-sensitive Hypothesis Verification

A common phenomena found in different fields of natural processing, such as automatic speech recognition, information retrieval or question answering, is that processing techniques seem to hit a ceiling of performance. In terms of contextual computing, automatic speech recognition systems employ by and large two knowledge sources:

- acoustic models in the form of Hidden Markov Models (HHM), which are learned from previously recorded data using either maximum likelihood criteria (with the Baum-Welch algorithm or gradient-based methods) or maximum mutual information criteria (with gradients respecting transition or observation probabilities);

- language models in the form of statistical n-gram models derived from previously collected and transcribed data.

Some low-level context-dependent features have been added to handle dialects and speaker-adaptation, dynamic lexica [Rapp et al., 2000] and recently out-of-vocabulary recognition techniques to enable the recognizer to detect unknown words and switch to pure phoneme recognition [Fetter, 1998]. However, no explicit contextual knowledge, e.g. of the domain or situation at hand is

Table 3.2: Domain Context - The Tasks T_A T_B T_C

Task Name	Task description
T_A (Accurate)	classification of the correctness of each hypothesis as a representation of the user's intention into correct versus incorrect based on the domain context.
T_B (BestOf)	choosing the best speech recognition hypothesis from an n-best list of hypotheses based on the domain context
T_C (Coherence)	classification of the coherence of each hypothesis into coherent versus incoherent based on the domain context

taken into account, which leaves the problem of dealing with phonetically indistinguishable input, unresolved and consequently produces noise. The classic example in the community is, that a large vocabulary speech recognition (LVSR) system, as needed for more conversational dialogue systems, could hardly differentiate between homonymic utterances such as: "*it is hard to wreck a nice beach*" and "*it is hard to recognize speech*". Humans on the other hand *hear* either one or the other depending on the context.

Noise in Speech Recognition: N-best Lists

Today's LVSR systems rarely feature simple one-best hypothesis as interface between ASR and NLU. While that may suffice for restricted dialogue systems, most systems either operate on n-best-lists as ASR output or convert ASR word graphs [Oerder and Ney, 1993] into n-best lists, given the distribution of acoustic and language model scores [Schwartz and Chow, 1990].

In our data a user expressed in example 13 the wish to see a specific city map again, leading to the top two speech recognition hypotheses (examples 14 and 15). In the annotators experiment described below, annotators found that example 14 constituted a more plausible representation of the utterance whereas example 15 constituted a less adequate representation thereof:

(13) Ich würde die Karte gerne wiedersehen
I would the map like to see again

(14) Ich würde die Karte eine wieder sehen
I would the map one again see

(15) Ich würde die Karte eine Wiedersehen
I would the map one Good Bye

Facing multiple representations of a single utterance consequently poses the question which of the different hypotheses most likely corresponds to the user's utterance. Several ways of solving this problem have been proposed and implemented in various systems, i.e.

- to use scores provides by the ASR system, i.e. acoustic and language model probabilities [Schwartz and Chow, 1990]; or

- to use scores provided by the natural language understanding and discourse modeling components [Litman et al., 1999a, Pfleger et al., 2002].

Consequently, in this work, the research question to be addressed below is how additional contextual knowledge - i.e., knowledge of the domain - can be employed beneficially for enhancing a dialog system's performance on this task. In the following I will, therefore, report on the experimental setup and evaluations of this question, thereby introducing the notion of *contextual coherence*. As described in Section 3.1, models of domain knowledge fill the *model column* for domain context in the context categories presented in Table 2.7 and have been employed in natural language understanding systems [Allen et al., 1995, Ferguson and Allen, 1998]. The scoring procedure described herein can be employed independently of the specific ontology modeling language used, as the underlying algorithm operates only on the nodes and named edges of the directed graph represented by the model. This specific domain model is, then, converted into a graph, consisting of:

- the class hierarchy, with each class corresponding to a concept representing either an entity or a process;
- the slots, i.e. the named edges of the graph corresponding to the class properties, constraints and restrictions.

3.2.2 The Data: Collection & Annotation

A first step towards solving classification problems in the area of human language processing is to test the hypothesis that humans are able to solve the classification problems reliably according to the predefined classification scheme. This classification scheme should ideally be determined by the concrete *tagging* task of the classification system. As a precursor step before this, a clear definition of the given tasks needs to be given, which can also be translated into an annotation scheme for human mark-up As such it needs to provide a set of disjunct levels of annotations for the individual discriminatory decisions that can be performed on spoken dialogue data, ranging from annotating referring expressions, e.g., named entities and their relations, anaphora and their antecedents, to word senses and dialogue acts. Each task must, therefore, have a clearly defined set of markables, attributes and values for each corpus of spoken dialogue data.

Methodology We will employ a uniform and generic method for computing task-specific sets of markables, sets of values and baselines for a given task T_w from the entire set of task, i.e. $T = \{T_1, \ldots, T_z\}$ and $T_w \in T$. I will present the conspecific figures of the markables, sets of values and baselines of our first tasks $T = \{T_A, T_B, T_C\}$ in Table 3.6. A gold standard annotation of a task features a finite set of markable tokens $W = \{w_1, \ldots, w_n\}$ for task T_w, e.g. if $n = 2$ in a corpus containing only the two ambiguous lexemes *bank* and *run* as markables, i.e. w_1 and w_2 respectively. For a member w_i of the set C I can now define the number of values for the tagging attribute of `sense` as: $A_i = \{b^i_1, \ldots, b^i_{n_i}\}$. For

Table 3.3: Domain Corpus - SRH_0

Corpus Name	SRH_0
Data Collection	Hidden Operator Test
Subjects	29
User Utterances Annotated - turns	1479
Speech Recognition Hypotheses - SRHs	2300
SRH per turn	1.55

example, for three senses of the markable *bank* as w_1 the corresponding value set is $A_1 = \{\texttt{building}, \texttt{institution}, \texttt{shore}\}$ and for *run* as w_2 the value set $A_2 = \{\texttt{motion}, \texttt{storm}\}$. Note that the value sets have markable-dependent sizes.

Data Collection

In order to craft annotation schemes, test the reliability of such annotations and compute baselines a corpus of spoken language data was employed. The data collection was conducted by means of a hidden operator test, which was designed as a light-weight Wizard-of-Oz experiment [Rapp and Strube, 2002]. In the test the system was simulated. Altogether 29 subjects were prompted to say certain inputs in 8 dialogues. 1479 turns were recorded. Each user-turn in the dialogue corresponded to a single intention, e.g. an instructional request (for example *How do I get to the Peterskirche*) or informational request (for example *Give me information about the Peterskirche*). The collected audio files were sent to the speech recognizer and the resulting n-best lists of SRH, were recorded in log-files. The final corpus consisted of approximately 2300 hypotheses. This corresponds to approx. 1.55 speech recognition hypotheses per user's turn.

An additional corpus of 1375 hypotheses was obtained by generating a new unseen set of SRH using the final SmartKom system. The collected n-best lists of SRH, were recorded in log-files. The final corpus consisted of 552 utterances. This corresponds to approximately 2.5 speech recognition hypotheses per user's turn. The data obtained from the hidden operator and system tests had to be processed to compose a corpus with n-best speech recognition hypotheses. For this purpose, the files were converted into the audio format of the SmartKom system and fed to the speech recognition component [Berton et al., 2006]. The input for the domain modeling component, i.e. n-best lists of speech recognition hypotheses were recorded in log-files and then processed by a set of conversion scripts. The speech recognition hypothesis corpus was then transformed into a set of annotation files which could be read into MMAX, an annotation tool adopted for this task [Müller and Strube, 2001]. Though originally developed for annotating other phenomena, the tool is equally suitable for this task, given an annotation scheme it can also compute statistics for the reliability of annotations.

Table 3.4: Domain Corpus - SRH_1

Corpus Name	SRH_1
Data Collection	Hidden Operator Test
Subjects	29
User Utterances Annotated - turns	552
Speech Recognition Hypotheses - SRHs	1375
SRH per turn	2.5

The Data Annotation & Reliability

In any annotation experiment first of all the central question can be stated in the terminology given above, as whether it is possible to reliably annotate a given corpus, here one of speech recognition hypotheses, with a given set of values. The motivation for that, as mentioned in Section 3.2.2, was to find out whether it is feasible to examine the contribution of domain knowledge to contextual computing on the task of classifying speech recognition hypotheses using the values reflected in the annotation scheme. I will, therefore, firstly present agreement measures for human annotators tagging the attribute *contextual coherence* with the two values *coherent* and *incoherent*.

This discrimination task was designed to exclude other contextual influences, e.g. discourse context (added and discussed in section 3.3 was solely based on general world-knowledge. I will, subsequently, show the feasibility of determining internal coherence of the output of the speech recognizer. Based on these results - i.e. human annotators classifying 2300 speech recognition hypotheses reliably in terms of their domain-specific coherence, subsequent experiments on coherence measures, best-SRH classification, word sense disambiguation and relation extraction ensued which will be discussed in Sections 3.3 through Section 3.6.

In an initial annotation experiment, the first data set of hypotheses were randomly mixed to avoid any priming influences and given to separate annotators for classifying each SRH as a markable with the values of coherent versus incoherent. To measure the reliability of annotations the so-called *kappa statistic* is frequently employed - where applicable - as the overall coefficient of agreement between annotators [Cohen, 1960, Carletta, 1996]. For this annotation tools can facilitate the automatic computation of the kappa metric on annotated files [Poesio and Vieira, 1998, Müller and Strube, 2001]. The resulting kappa statistic, employing Formula 2.4, over the annotated data were K=0.7 given this first annotation of coherent versus incoherent [Gurevych et al., 2002], which clearly suffices to say that human annotators can quite reliably differentiate between coherent samples (as in Example (14)) and incoherent (as in Example (15)). In this experiment 1479 utterances from the dialogues collected in the hidden-operator tests - where each utterance corresponded to a single intention, e.g. a route- or a sight information request - were used as the initial set of transcrip-

Table 3.5: Task Coherence - Annotation Experiment - SRH_0

Data Source	Corpus SRH_0
User Utterances Annotated - turns	1479
Speech Recognition Hypotheses - SRHs	2300
Classification Values	coherent - incoherent
Agreement: Kappa	$\kappa = 0.7$

tions and the corresponding speech recognition output, i.e. n-best lists of SRHs for all utterances. This initial corpus is overviewed in Table 3.3.

Human Annotations for the Evaluation Tasks A, B and C

For the ensuing experiments a new subset of the hidden operator corpus was used for second set of annotation experiments as well as an evaluation of the system. For this the scores of the recognizer and other components of the system were logged next to the contextual coherence scores discussed below. This trial resulted in a sub-corpus of 552 utterances corresponding to 1.375 SRHs along with the respective confidence scores; described in Table 3.4. The data resulting from the new set of annotation experiments were employed to produce a hand-annotated corpus to be used as a so-called *gold standard* for the evaluation of the contextual coherence scores. Furthermore, given new annotation tasks one should test anew whether human subjects are able to annotate the data reliably according to the novel annotation schemata. Therefore, two annotators were specially trained for each of these particular annotation tasks.

Note that in the initial annotation experiment, the task of annotators was to classify a subset of the corpus of SRHs as either coherent or incoherent where the hypotheses were randomly mixed in order to avoid contextual priming and a metric, such as kappa was applicable to express inter-annotator agreement as shown in Table 3.5. In the subsequent annotation experiments the markables were presented in their dialogical order together with the human transcription in case of Task A and B and without it for Task C, where the hypothesis' internal coherence was to be determined regardless of the actual utterance that caused it. In order to present and compare performance measures - as discussed in Section 2.5 - and baseline measures - as discussed in Section 2.6 in a uniform way - I will express the performance of human annotation, automatic classification and baseline approach in terms of precision on the given classification task. This corresponds to p in Formula 2.2 or a corresponding f-measure with $\alpha = 1$ as discussed in Section 2.5.3.

Task A: Annotator Performance

Given the transcribed corpus of utterances and their corresponding speech recognition hypotheses SRH_1, the underlying question of this annotation experiment

Table 3.6: Classification Values and Annotation Performance for Corpus SRH_1

Data Source Corpus SRH_1		
Task	Classification Values	Human Performance
T_A	correct - incorrect	precision $\approx .8$
T_B	best - non-best	precision $\approx .95$
T_C	coherent - non-coherent	precision $\approx .8$

was if the amount of *noise* produced by the speech recognition system distorted the *signal* to a degree that the correct intention behind it can no longer be recovered.[3] Such hypothesis are, consequently, to be annotated with the value *incorrect*. In case the amount of noise is tolerable and the intended meaning is still correctly represented, the corresponding annotation value was *correct*.

In the experiment for Task A - *Accurate* - the results and values are displayed in Table 3.6. They show that annotators could reliably identify accuracy in *overall meaning*, i.e. differentiate between correct or incorrect representations of the transcribed utterance and the corresponding SRH. Given the corpus SRH_1, the annotators reached an agreement of 80% given the task to classify the correctness of each hypothesis as an accurate representation of the user's intention. Clad in terms of performance, they reached a precision of $\approx .8$ relative to each other.

Task B: Annotator Performance

In the experiment for Task B - BestOf - the values and results are also displayed in Table 3.6. They show that annotators could also reliably identify the best hypothesis, given a transcribed utterance and the corresponding SRHs generated by the speech recognition system. Given the corpus SRH_1, the annotators reached an agreement of 95.35%. Also in this experiment, the annotators saw the SRHs together with the transcribed user utterances. The task of annotators was to determine the best hypothesis from the n-best list of SRHs corresponding to a single user utterance. The decision had to be made on the basis of several criteria.

The most important criteria was how well the SRH captures the intentional content of the user's utterance. If none of the SRHs captured the user intend adequately, the decision had to be made by looking at the actual word error rate. In this experiment the inter-annotator agreement was 95.69% corresponding to an excellent relative performance of $\approx .96$

A kappa-statistic suitable for measuring the reliability of annotations is not applicable straight-forward in this case, as there are heterogeneous markable sets

[3] Please note that in this case *noise* corresponds to speech recognition errors, i.e. insertions, substitutions and deletions as discussed in Section 2.5.2, and the *signal* corresponds to the utterance seen as an ordered sequence of strings.

for each classification, due to the different number of SRHs in the n-best lists. However, for the chosen annotator-relative performance measure this does not constitute a problem as it is calculated based percentage of utterances, where the annotators agreed on the best SRH.

Task C: Annotator Performance

The results of the experiment on Task C - Coherence - replicate the findings of the initial experiment - described above - for the classification of coherence, i.e., that novel annotators could again reliably assign the values *coherent* and *incoherent*, given a transcribed utterance and the corresponding SRHs choices generated by the ASR system. Based on the new corpus SRH_1, the annotators reached an agreement of 79.71% or performed with a precision of $\approx .8$ relative to each other.

Next, I will present the results of employing first solely domain-specific knowledge for automatically classifying the corpus SRH_1 of recognition hypotheses using the formal model of the domain that was derived from different corpus of data [Schiel and Türk, 2006]. This approach can be employed by any language understanding system to classify utterances, e.g. n-best list thereof, as instances of a given domain model, e.g. the ontology described in Section 3. The corresponding algorithm will be described below, followed by its evaluation against the human annotation-based *gold standards* for Task A, B and C, where each task-specific doubly annotated corpus was transformed into an evaluation *gold standard* by means of the annotators agreeing on a single solution for the cases of disagreement. I will now describe the algorithm, its corresponding performance on the three tasks and their individual baselines metrics derived from the individual gold standards of the three annotation experiments described above [Porzel et al., 2003a].

3.2.3 The Algorithm: Domain-specific Coherence

The algorithm performs a number of processing steps and routines, each of them and the corresponding data structures will be described separately in the respective subsections.

Obtaining Input

A necessary preprocessing step is to convert each markable into a set of *concepts* that label nodes in the graph employed to represent the domain context. This can be achieved in three ways depending on the type of input:

- Unstructured input, e.g. a set of lexical strings. For this one employ a corresponding lexicon enhanced with specific concept mappings. That is, for each entry in the lexicon it can be marked as being an instance of none, one or many classes of the ontology. A set of the concepts, corresponding to the classes of which the lexical strings - in the original set of strings - are instances, constitutes the resulting input. All other strings

with no concept mappings, e.g. articles, are ignored in the conversion. More elaborate ontological models of lexica have been proposed that also include a detailed morpho-syntactic specification [LMF, 2005, MAF, 2005, SynAF, 2005]; also a fully integrated model within a ground ontology has been proposed at the cost of making the resulting ontological model to be expressible only in OWL-Full and higher [Buitelaar et al., 2006].

- Semi-structured input, e.g. some local *semantic* grammar exists, as found in the output of different parsers [Abney, 1996, Pinkal et al., 2000], production systems [Engel, 2002] or semantic analyzers [Bryant, 2003] or some legacy data model of an information extraction system [Pivk et al., 2006].

- Structured input, e.g. when the system is ontology-based so that all instances employed in communication are based on one ontological grammar and ontological vocabulary, as implemented in the SmartWeb system [Reithinger et al., 2005, Ankolekar et al., 2006, Oberle et al., 2007].[4]

In all cases ambiguities may arise, e.g. in case of multiple mappings from a lexical item to several nodes in the domain model. The algorithm, to be presented below, regards each enumeration of the possible mappings from the recognized set of strings at hand individually and selects the most coherent enumerations score - to be defined below - as the score for the source recognition hypothesis. The problem of ambiguity, and the algorithm's performance therein, will be discussed in greater detail in Section 3.4, after presenting the way in which domain knowledge can be employed for contextual computing.

The Scoring Algorithm

First of all, the algorithm converts the domain model, i.e. the ontology, into a directed graph with concepts as nodes and relations as edges. In order to enable the algorithms to ascend the class hierarchy upwards as well as downwards the graph was enriched during its conversion by *parent-of* relations symmetric to the *subclass-of* relation. This eliminates directionality problems as well as avoids cycles and 0-paths. In order to find the shortest path between two concepts, the algorithm employs the *single source shortest path* algorithm - also called the *Dijkstra* algorithm [Dijkstra, 1959, Cormen et al., 1990].

Given a set of concepts C $\{c_1, ..., c_n\}$, the algorithm runs once for each concept. The Dijkstra algorithm calculates minimal paths from a source node to all other nodes. Then, the minimal paths connecting a given concept c_i with

[4] In case of the SmartKom system the communication was not based on an RDF-schema model, but restricted to XML documents syntactically specified by XML-schemata, which nonetheless enables the parser and NLU system's output to be defined in the corresponding ontological vocabulary, e.g. by employing an automatic approach to specify the schema interface specifications - on the XML-level - based on a formal RDFS model. The model that has been used for the SmartKom system was already described in Section 3.1.3 above. Due to the restrictions of XML schema, this approach requires a strictly reductionistic ontology, i.e. one without multiple inheritance [Gurevych et al., 2003a].

every other concept in C (excluding c_i itself) are selected, resulting in an $n \times n$ matrix of the respective paths.

To score the minimal paths connecting all concepts with each other in a given set C, the algorithm employed a approach that singled out non-taxonomic relations - to be regarded as the *semantic* edges of nodes modeled in an non-semantic *isa* hierarchy [Demetriou and Atwell, 1994]. In this approach, $R = \{r_1, r_2, \ldots, r_n\}$ is the set of direct relations (both *isa* and *semantic* relations) that can connect two nodes (concepts); and $W = \{w_1, w_2, \ldots, w_n\}$ is the set of corresponding weights, where the weight of each *isa* relation is set to 0 and that of each other relation to 1. For each two concepts c_i, c_j the set $P = \{p_1, p_2, \ldots, p_m\}$ denotes the scores of all possible paths that link the two concepts. The score for path $k(k = 1, \ldots, m)$ can be given as shown in Formula 3.1.

$$p_k = \sum_{i=1}^{n} a_i w_i \tag{3.1}$$

where a_i represents the number of times the relation r_i exists in path k. The ensuing distance between two concepts c_i and c_j is, then, defined as the minimum score derived between c_i and c_j, as shown in Formula 3.2.

$$D(c_i, c_j) = min(p_k) \quad k = 1, 2, \ldots, m \tag{3.2}$$

The algorithm then selects from the set of all paths between two concepts the one with the smallest weight, i.e. the *cheapest*. The distances between all concept pairs in C are summed up to a total score. The set of concepts with the lowest aggregate score represents the combination with the highest semantic relatedness.

Since the objective is to compute a coherence score based on a domain model for given arbitrary sets of concepts that are part of the vocabulary of the model on a specific scale, additional extensions are necessary. In this experiment, the concept sets to be scored can differ in terms of their content, the number and their mappings from the original speech recognition hypothesis. Moreover, the final score could reflect the number of concepts in an individual set given the number of lexical items in the original hypothesis. Additionally, the results must be normalized in order to allow for evaluation, comparability and clearer interpretation of the semantic coherence scores.

A Domain-specific Score for Sets of Concepts

In order to make the algorithm described above applicable and evaluatable with respect to the task at hand as well as other possible tasks a maximal distance between two concepts c_i and c_j that are only connected via *isa* relations in the model needs to be determined as D_{max}. The basic idea is to calculate a score based on the semantic distances in the set C and to let short distances indicate coherence and a greater distance for concept pairs in a given C that have no *semantic path*. A maximum value can serve as a cut-off for long distances

and, thus, help to prune the search tree for long and semantically irrelevant (redundant) transversals of the taxonomy. This constant has to be set according to the structure of the model. For example, employing the ontology described in Section 3.1.3, the maximum distance between two concepts does not exceed ten and D_{max} can therefore be set accordingly.

A domain coherence score can now defined a set C as the average path length between all concept pairs in C as shown in Formula 3.3.

$$S(C) = \frac{\sum_{c_i,c_j \in C, c_i \neq c_j} D(c_i, c_j)}{|C|^2 - |C|} \tag{3.3}$$

Since the ontology is a directed graph, there are $|C|^2 - |C|$ pairs of concepts with possible directed connections, i.e., a path from concept c_i to concept c_j may be completely different to that from c_j to c_i or even be missing. As a symmetric alternative, one may want to consider a path from c_i to c_j and a path from c_j to c_i to be semantically equivalent and thus model every relation in a bidirectional way. In that case a symmetric score $S'(C)$ can be computed as given in Formula 3.4.

$$S'(C) = 2\frac{\sum_{c_i,c_j \in C, i<j} min(D(c_i, c_j), (D(c_j, c_i))}{|C|^2 - |C|} \tag{3.4}$$

The algorithm implemented both options: one for domain models that feature more bi-directional relations (via axiomatization or explicit modeling) and one for domain models that feature more uni-directional relations.

In the ontology used for this experiment some bi-directional relations can be found, e.g. given c_1=*Broadcast* and c_2=*Channel*, there exists a path from c_1 to c_2 via the relation *has-channel* and a different path from c_2 to c_1 via the relation *has-broadcast*. However, such reverse relations are only sporadically represented in the ontology. Consequently, it is difficult to account for their influence on $S(C)$ in general. That is why we chose the $S'(C)$ function for the evaluation, i.e. only the best path $D(c_i, c_j)$ between a given pair of concepts, regardless of the direction, is taken into account.

In order to score the alternative nodes from the domain model - defined by $I'(input_{n+1})$ - the function shown in Formula 3.4 is employed. This means a domain context coherence score S' is calculated for each domain-specific concept set C'. To let a higher number indicate more domain-specific coherence an inverse linear transformation of the scores is performed resulting in numbers from 0 to 1.

Given the algorithm presented above, a significant number of misclassifications for SRHs would result from the cases when an SRH contains a high proportion of function words (having no conceptual mappings in the resulting concept set C) and only a few content words which are defined as instances of a class in the domain model. To illustrate consider the set of strings given in Example 16.

(16) Wo den Informationen zu das gleiche
Where the information to the same

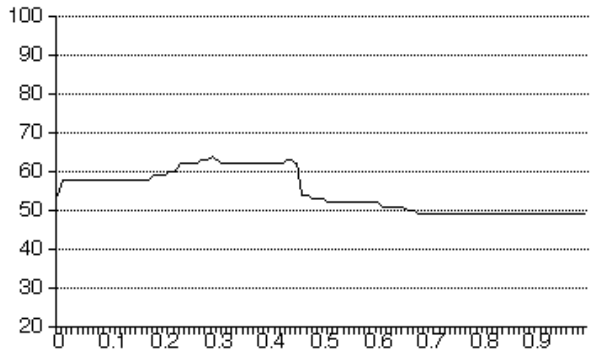

Figure 3.4: Determining the optimal threshold on the coherent *versus* incoherent classification data from corpus SRH_0. The vertical axis shows performance and the horizontal shows the word to concept threshold V

Depending on the given lexical mappings the corresponding concept set could be constituted out of a single concept, e.g. one concept mapped to *Information*. This would classify the set as coherent with the highest possible score, as this is the only concept in the set. This, however, would often lead to misclassifications. To eliminate this effect and to discard linguistically slighted distributions of function and content words a post-processing technique takes the relation between the number of ontology concepts N_c in a given concept set and the total number of words N_w in the original SRH into account. This relation is defined by the ratio $V = N_c/N_w$. Therewith an automatic classification of an hypothesis as being incoherent - irrespective of its coherence score $S(C)$, by setting a threshold V. The threshold may be set freely. Employing the gold standard data an optimal threshold can be set once for each task. In all cases this threshold is found where it would be linguistically expected, i.e. where the set contains at least one concept per three lexical forms found in the input, as exemplified in the dependency graph of the threshold value V and the results in performance for the task of coherence using corpus SRH_0 is shown in Figure 3.4. In the findings presented in the evaluation section below the resulting threshold values employed were $V = .33$ for Task A, $V = .34$ for Task B and $V = .39$ for Task C.

Looking at an example of the algorithm at work, I will employ the utterance given in Example 13[5] The resulting two SRHs - SRH_1 and SRH_2 - are given in Examples 14 and 15 respectively. The human annotators considered SRH_1 to be coherent and labeled SRH_2 as incoherent. Mapping the contained lexical instance to concepts, each SRH is transformed into a distinct concept sets as shown in Table 3.7. As no ambiguous words are found in this example, C_1

[5]Example 13 was *Ich würde die Karte gerne wiedersehen* or be glossed as *I would the map like to see again* in English.

Table 3.7: Example concept sets and labels

Concept Set	c_1	c_2	c_3
C_1	Person	Map	Watch Perceptual Process
C_2	Person	Map	Parting Process

Table 3.8: Example semantic and taxonomic paths and distances

Set	Distance	Property Definition
C_1	1	has-watcher(Watch Perceptual Process, Person)
C_1	1	has-watchable_object(Watch Perceptual Process, Map)
C_1	10	via *isa* relations(Person, Map)
C_2	1	has-agent(Parting Process, Person)
C_2	10	via *isa* relations(Person, Map)
C_2	10	via *isa* relations(Parting Process, Map)

corresponds to SRH_1 and C_2 corresponds to SRH_2
Once each hypothesis is mapped unto the graph the algorithm determines all paths between the concepts of each set, of which only the semantic edges are correspondingly weighted for the scoring. This yields the following semantic paths for C_1 the domain models states a) that a person can be the watcher of a perceptual watching process and b) that a map can be a watchable object of such a process as shown in Table 3.8.
The ensuing average distance between the concepts of C_1 and C_2 using S' - given in Formula 3.4 - is:

total distance	number of concept pairs	average using S'
$(C_1) = 12$	3	$S'(C_1) = 4$
$(C_2) = 21$	3	$S'(C_2) = 7$

The corresponding result for S - given in Formula 3.3 - is:

total distance	number of concept pairs	average using S
$(C_1) = 42$	6	$S(C_1) = 7$
$(C_2) = 51$	6	$S(C_2) = 8.5$

In both cases the results are sufficient for a relative judgment, i.e. SRH_2 constitutes a less semantically coherent structure than SRH_1. To allow for a binary classification of each hypothesis as either coherent or incoherent as well as either correct or incorrect, a cutoff score must be set unlike in Task B (Best-of), where the best scoring hypothesis wins. The settings employed in the experiments of Task A and C will be presented shortly along with their corresponding results in Section 3.2.4.

Please note that, given the results presented below, a corresponding domain-specific contextual computing component has been integrated in the SmartKom prototype multimodal dialog system [Porzel et al., 2006a] and provides a coherence score for each hypothesis generated in the speech processing pipeline of the system. In this implementation it is employed by the natural language system to determine the best hypothesis from the n-best lists in conjunction with acoustic or statistical scores provided by other speech and language processing component. Next to the application of the domain-specific scoring algorithm the SmartKom system also employs the ontology described in Section 3.1.3, which is also used in this experimental setting. In the following, the evaluation of the performance of the classification system will be presented below in the light of the human *gold standard* and a set of baseline measures computed for Tasks A, B, and C.

3.2.4 Results: Domain-sensitive Hypothesis Verification

For this evaluation of algorithm the dataset of corpus SRH_1 - presented in Table 3.4 was employed. Based on the double human annotations - presented in Table 3.6 - for each classification task, three corresponding *gold standards* were crafted via inter-annotator negotiations. Given these gold standard mark-up for the Tasks Accurate (T_A), BestOf (T_B) and Coherence (T_C), a corpus-based computation of a baseline metric has additionally become feasible, as I will describe below and apply for a discussion of the evaluations results thereafter.

Methodological Baseline Computation

Baselines for classification or tagging tasks have been discussed in Section 2.6.2. The baselines presented herein are corpus-based using the individual gold standards to compute a majority class performance on this annotated data. Thus, I can, again, provide the corresponding f-measure with $\alpha = 1$ as expressed in Formula 2.2, which I have given for measuring human and algorithmic performance. The baseline performance, therefore, mirrors an analogous evaluation of a component that always chooses the most frequent solution - given herein as majority class performance metric for a given task against the same gold standard.

A gold standard annotation of a task features a finite set of markable tokens $W = \{w_1, \ldots, w_n\}$ for task T_w, e.g. if $n = 3$ in a corpus containing the ambiguous lexeme *bank* as in our initial Examples 1, 2 and 3, as markable, i.e. w_1. For a member w_i of the set W I can now define the number of values for the tagging attribute of `sense` as: $A_i = \{b^i_1, \ldots, b^i_{n_i}\}$. For example, for three senses of the markable *bank* as w_1 we get the corresponding value set $A_1 = \{$`database`, `building`, `shore`$\}$ for Examples 1 through 3. Note that the value sets can have markable-dependent sizes. For computing the proportional majority classes one need to compute the occurrences of a value j for a markable i in a given gold standard test data set, called V_{ij} herein. Now, it is possible to determine the

most frequently given value and its number for each markable c_i as shown in Formula 3.5.

$$V_i^{max} = max_{ij \in \{1,\ldots,b_i\}} V_{ij} \tag{3.5}$$

The total number of values for a markable c_i can be defined as in Formula 3.6.

$$V_i^S = \sum_{j=1}^{n_1} V_{ij} \tag{3.6}$$

In Formula 3.7 V_i^{max} is defined as the majority class baseline.

$$B_i = \frac{V_i^{max}}{V_i^S} \tag{3.7}$$

If an classification algorithm always chooses the most frequent attribute for markable c_i, the likelihood of correct guesses corresponds to B_i. The total number of values can, consequently, be calculated as shown in Formula 3.8.

$$V^S = \sum_{i=1}^{n} V_i^S \tag{3.8}$$

Based on this it is possible to compute the task-specific proportional baseline, given in Formula 3.9, for task T_w, i.e., B_{T_w}, over the entire test set.

$$B_{T_w} = \frac{1}{V^S} \cdot \sum_{i=1}^{n} V_i^S B_i = \frac{1}{V^S} \cdot \sum_{i=1}^{n} V_i^{max} \tag{3.9}$$

Thus, B_{T_w} calculates the average of correct guesses for the majority baseline. Additionally, one also compute different individual majority class baselines for each markable as well as a total number of values given for all w_n out of W.

A lower measure on this baseline metric of one corpus as compared to another indicates that the method of choosing for each markable always the most frequently occurring class would perform worse on the former corpus then the latter one. This being the basic desideratum of the any baseline metric, I will note that proportional baseline measure is also able to compute the performance of such a majority class-based approach on any data set for any classification task - as will be further exemplified in Section 3.4.4 - before presenting the baseline metrics for Task A, B, and C.

Determining the Task-specific Majority Class Baselines

In the following, I will present the individual majority class baselines computed for Tasks A, B and C. In Section 3.2.2, I have already presented task-specific relative human annotation performances, which - in terms of performance reliability - can be regarded as an upper bound baseline measure, whereas the respective majority class baselines constitute an lower bound one.

Task A: Baseline Performance

A majority class baselines baseline for the evaluation of algorithm on Task A was derived by considering each speech recognition hypothesis a markable in the set $W = \{w_1, \ldots, w_n\}$ which contained a total number of 1375 given the corpus SRH_1 at hand. The values of the attribute *Accurate* were *correct* and *incorrect*. In the gold standard employed for the evaluation of algorithm - which was derived from the annotation experiment described in Section 3.2.2 - 51.93% of all markables were classified as *correct*. This constituted the majority class for this corpus and consequently leads to a majority class baseline performance of $\approx .52$ as shown in Table 3.9

Task B: Baseline Performance

Again, each speech recognition hypothesis was considered as a markable in the same set W of corpus SRH_1. The values of the attribute *BestOf* were *best* and *non-best*. In the case of this task, keeping the set of markables W constant to Task A as well as to Task C corresponds to the classification conditions for the respective annotation tasks. While having the same markable set in all annotations, baseline metrics and evaluations brings about certain advantages, the consequential drawback in case of Task B is that - in cases where in the recognition system produced only one hypothesis for a given utterance - the classification task to assign the value *best* becomes trivial. To treat all experiments alike in this respect also gave a corresponding advantage to the human annotators - as reflected in their high performance relative to the one obtained in the other tasks. Besides being fair-handed, the evaluation results to be presented below, show that the mean distance between the algorithm's performance on the three task and that of the relative human performance was .12 with a variance of $\pm.03$ while the average human performance on the tasks varied by more than threefold. Given the gold standard markable set employed for the all evaluations of the algorithm 63.91% of all markables were classified as *best* by the human annotators. This constituted the majority class for this corpus and consequently leads to a majority class baseline performance of $\approx$.64% as shown in Table 3.9.

Task C: Baseline Performance

A majority class baselines baseline for the evaluation of algorithm on Task C was derived by, again, considering each speech recognition hypothesis in corpus SRH_1 as a markable in the set W. The values of the attribute *Coherent* were *coherent* and *incoherent*. In the gold standard employed for the evaluation of algorithm - again derived from the annotation experiment described in Section 3.2.2 - 63.05% of all markables were classified as *correct*. This constituted the majority class for this corpus and consequently leads to a majority class baseline performance of $\approx$.63% as shown in Table 3.9.

Table 3.9: Classification Values and Baseline Performance for Corpus SRH_1

Data Source Corpus SRH_1		
Task	Classification Values	Majority Class Baseline Precision
T_A	correct - incorrect	$\approx .52$
T_B	best - non-best	$\approx .64$
T_C	coherent - non-coherent	$\approx .61$

.

Classification Results

The experimental results presented below employ a classification system that - once an inverse linear transformation of the scores produced by S', given in Formula 3.4 - which there range from 1 to D_{max}) - is performed the output produced is a score on the scale from 0 to 1. Now, higher scores reflect a lower average semantic path-length of the shortest graph found to connect the concepts in the set C, given the ontology described in Section 3.1.3. As discussed in Section 3.2.3, all sets where V was smaller than $V = .33$ for Task A, $V = .34$ for Task B and $V = .39$ for Task C - as a result of dividing the number of lexical items in the SRH by the number of concepts in C - were set to D_{max} and now become 0.

Task A: Classification Performance

For the case of Task A (and also C) a score had to be determined for making a binary value distinction, as in the case of V - discussed in Section 3.2.3 - an optimal score was found for Task A by going about half the way towards D_{max} at 4.2. This means that all SRHs found below .45 on the inversely linear transformed score where subsequently classified as incorrect and those above as correct in this classification experiment.

Employing the classification system described above, the experimental classification yields the precision of $\approx .65$ on task T_A. This means that in 65.09% of all cases a markable - a speech recognition hypothesis - defined by the gold standard as correct is classified as such by the classification system. Which constitutes a gain of $\approx .13$ points over the corresponding baseline of $\approx .52$, which only classified 51.93% - or 8.16% less - of the gold standard markables correctly.

It, therefore, also lies $\approx .15$ points below the human relative performance of $\approx .8$. - this direct comparison, however, has to be taken with a grain of salt as the difference of 14.91% subtracts the 65.09% correctly classified markable from of the 80% correctly classified markables of an annotator - employing the alternate annotator's mark-up as a gold standard - which constitutes a different gold standard then the merged one employed for determining the algorithm's percentage. These results are presented together with the results of the human reliability and majority class baseline metrics in Table 3.10.

Task B: Classification Performance

For the experimental setting of Task B no score for making a binary value distinction is needed, as the highest scoring hypothesis from each set was classified *best* and the rest correspondingly as *non-best*. As stated above, the word to concept threshold, V, was set to .34. Employing the ensuing classification system, the experimental classification yields the precision of $\approx .84$ on task T_B. This means that in 84.06% all the markables defined by the gold standard as *best* are classified as such by the classification system.

This result constitutes a gain of $\approx .2$ points over the corresponding baseline of $\approx .64$, which only classified 64% - or 20.06% less - of the gold standard markables correctly. Looking at the human relative performance of $\approx .95$ this result lies still $\approx .11$ points below - or 10.94% as the difference between 95% and 84.06% correctly classified. These results are also presented together with the results of the human reliability and majority class baseline metrics in Table 3.10.

Task C: Classification Performance

As in the case of Task A a score for achieving a binary value distinction is also needed in Task C. As in the case of the threshold V an optimal score was determined empirically. As it is quite possible for an speech recognition hypothesis to be marked incorrect in Task A, but still to be internally coherent in Task C, but not *vice versa* - to be both incoherent, but also a correct representation of the underlying intention - it follows that a higher number of markables labeled correct is to be expected as compared to ones labeled coherent.

This state of affairs is both reflected by the higher majority class baseline in Task C compared to Task A as well as in the higher permissiveness of the classification cut-off point at 7.1. This means that all SRHs found below .29 on the inversely linear transformed score were subsequently classified as incoherent and those above as coherent in this classification experiment. The threshold V, however, is less affected by this and can be found at 3.9.

Employing these settings, the experimental classification yields the precision of $\approx .7$ on task T_C. This means that in 70.4% of all cases a markable - a speech recognition hypothesis - defined by the gold standard as coherent is classified as coherent by the classification system. Which constitutes a gain of $\approx .07$ points over the corresponding baseline of $\approx .63$, which only classified 63.01 % - or 7.35% less - of the gold standard markables correctly. It, therefore, lies $\approx .1$ point below the human relative performance of $\approx .8$ established on a different gold standard for the corpus SRH_1. These results are presented together with the results of the human reliability and majority class baseline metrics in Table 3.10.

A statistical analysis and further discussion of the results presented above and summarized in Table 3.10 are given in Section 3.7. For the time being, I will note that the question behind classification tasks *accurate* and *coherent* is of a more basic nature, i.e. concerning the feasibility of separating correct and coherent from incorrect and incoherent hypotheses by employing domain context. In the case of the Task B (BestOf), there is also a straight-forward application

Table 3.10: Results of Classification Experiments for Tasks A, B and C

classification task on corpus SRH_1	majority class baseline precision	gain	classification algorithm precision	loss	relative human precision
Accurate	≈ .52	≈ .13	≈ .65	≈ .15	≈ .80
BestOf	≈ .64	≈ .20	≈ .84	≈ .11	≈ .95
Coherent	≈ .63	≈ .07	≈ .70	≈ .10	≈ .80

for employing domain context in the standard task of hypothesis verification, where the challenge lies in picking the best hypothesis out of the given n-best list. The performance of the classification system presented in this and the following Sections - with regard to the task of determining the best representation of an user's utterance - will also be discussed further in Section 3.3.4. Before doing so, I will turn to the question if the addition of another context type, i.e. discourse context, can contribute additional information that provides yet more pertinent knowledge about the overall context.

3.2.5 Roadmap

In order to continue this examination of the specific contributions of contextual computing in a set of tasks from the area of natural language processing - where the challenge lies in determining the most plausible item out of a set of possible alternatives as described in Section 2.7.2. The results of employing a specific domain model to score the possible alternatives have been presented above and - in the light of choosing a best-fitting speech recognition hypothesis out of a *noisy* set as in the case of Task B - show that this inclusion of domain knowledge as additional context classifies 20% more of the data correctly than a - hindsight-based and, therefore, somewhat informed - majority-class baseline. As discussed in Section 3.1.2, real utterances actually occur in the specific discourse context of what has been said before. In addition to considering each utterances by itself, the question of examining the potential contribution of adding domain concepts from the preceding discourse context to the set of discourse entities representing a speech recognition hypothesis - for which semantic paths are sought and scored - arises. I will present the corresponding algorithmic extensions and subsequent findings and in the following Section 3.3.

3.3 Using Discourse Context for Noisy Input

In a sense a given dialogical situation can be said to *evoke* a specific domain context in the background. The experiments presented so far have shown that interweaving corresponding domain knowledge, given as a formal and explicit model of a the domain at hand, brings about the specific gains and losses shown

in Table 3.10, by means of employing a semantic distance measure for arbitrary sets of nodes from that model. As discussed in Section 3.2.3 the situationally given discourse entities featured in the utterance, employing the terminology of Byron (2002), are mapped unto the vocabulary of the domain model. Once that mapping is performed a distance is computed considering the semantic paths connecting the individual concepts in the evoked domain model, as described in Section 3.2.3. In the examinations of contextual computing described so far this domain-specific algorithm was examined for a set of classification tasks on a corpus of noisy speech recognition hypotheses using an ontology as described in Section 3.1.3.

Next to the domain context evoked in a given dialogical situation, each utterance occurs in its own discourse context. I have presented models of this discourse context already along with their role and elementary functions in dialog systems in Section 3.1.2. Additionally, discourse context also constitutes the context type, which is employed numerously within dialog systems as well as in other computational approaches to discourse and text understanding on various levels of sophistication as discussed in Section 2.3 and Section 2.4.

Given the task described in the beginning of Section 3.2, it is now possible to asses the specific contribution of including discourse context on the three tasks for which performance results, task-specific gold standards and baseline metrics exist. I can, therefore present this addition of discourse context by applying the evaluation methodology employed in Section 3.2 to measure this contextual contribution to the individual classification tasks. Therefore, the necessary algorithmic extensions - creating a discourse sensitive way for scoring the individual discourse-enhanced sets of concepts - and task-specific performance measures will, consequently, be discussed in the following sections.

3.3.1 The Task: Discourse-sensitive Hypothesis Verification

In the following, I will describe the domain- and discourse-sensitive classification system and how it can be applied to estimate how well a given speech recognition hypothesis fits with respect to the existing models of domain and discourse context. Thereby, a second context type from the four types given in Table 2.7, i.e., discourse context, is added to domain context for the three hypothesis verification tasks listed in Table 3.2. The aim, therefore, remains to examine how context-sensitivity provides mechanisms to increase the robustness and reliability of dialogue systems. A consequential test is to also to examine if the discourse-sensitive algorithm can be employed by a spoken dialogue system to enhance the interface between automatic speech recognition and natural language understanding, which will be discussed further in Section 3.3.4.

3.3.2 The Data: Collection & Annotation

As stated above, it is possible in this examination to draw on identical lexical and ontological resources. Therefore, the markable set of speech recognition hy-

Table 3.11: Annotator & Baseline Performance T_A,T_B and T_C

Classification Task	Classifier	Precision on SRH_1
Task A	Human Annotators	$\approx .80$
(Accurate)	Baseline Majority Class	$\approx .63$
Task B	Human Annotators	$\approx .95$
(BestOf)	Baseline Majority Class	$\approx .64$
Task C	Human Annotators	$\approx .80$
(Coherent)	Baseline Majority Class	$\approx .52$

Table 3.12: Creating discourse-sensitive concept sets

$I(input_{n+1})$			$I'(input_{n+1})$
C_1	$\cup$	$C_{best}(input_n)$	$= C'_1$
C_2	$\cup$	$C_{best}(input_n)$	$= C'_2$
...		...	...
C_n	$\cup$	$C_{best}(input_n)$	$= C'_n$

pothesis from the SRH_1 corpus - described in Section 3.2.2 - can be employed. As for the classification task and the computation of the performance of human annotators and of the baseline method the same metrics will employed as above and discussed in Section 2.5. The annotated data, therefore, yield the same human- and baseline performances in terms of precision for the respective annotation experiments and gold standards in Table 3.11

3.3.3 The Algorithm: Scoring *cum* Discourse

A necessary preprocessing step for the discourse-sensitive concept scoring is to include discourse context into the concept representation to be valued as $C'(input_i))$ resulting from the following pair of concept sets:

- a concept set of the noisy input to be scored, i.e. $C(input_{n+1})$,

- and a concept set of the preceding input, i.e. $C(input_n)$.

For that purpose, the discourse-sensitive system stores the best concept representation from the preceding input as $C_{best}(input)$. The best set in this case is the one which received the highest score from the domain-specific system - described above - from the respective list of alternative representations for the utterance. To produce a conceptual discourse-enriched context set for $input_{n+1}$, a union can be build of each of its possible interpretations $I = \{C_1, C_2, \ldots, C_n\}$ with the stored $C_{best}(input_n)$ from the previous interpretations. This results in a contextually augmented new set $I' = \{C'_1, C'_2, \ldots, C'_n\}$ representing possible contextual concept interpretations of $input_{n+1}$.

If, however, the calculated score of $C_{best}(SRH_n)$ is below a certain threshold, meaning that even the best prior hypothesis is most likely not semantically coherent, then $C_{best}(SRH_n) = \{\emptyset\}$. Thusly, when $C_{best}(SRH_n)$ is empty only the concept sets for SRH_{n+1} are taken into account. This is, of course, also the case at the first dialogue turn.

In order to score the alternative discourse-sensitive context sets defined by $I'(SRH_{n+1})$, the scoring approach given in Formula 3.4 in Section 3.2.3 is employed. Thereby, one can calculate a domain- and discourse-context score S' for each conceptual context representation C'. Also, the same inverse linear transformation of the scores resulting in numbers from 0 to 1 can be performed, so that higher scores indicate better contextual coherence.

3.3.4 The Results: Discourse-sensitive Hypothesis Verification

Again, for the case of Task A and C a cut-off score had to be determined for making a classification between correct - equal or above the cut-off score - and incorrect - below the cut-off score in Task A as well as between coherent and incoherent in Task C. For this the same procedure can be applied as in the case of finding the optimal word-concept ratio V discussed in Section 3.2.3 and shown for the corresponding domain-specific classification tasks in Section 3.2.4.

Task A: Classification Performance

As described in Section 3.2.2, the task in this classification experiment on the corpus SRH_1 is to differentiate correct speech recognition hypotheses - which contain a tolerable amount of *noise* - from incorrect ones - where the users intention is too distorted to be deemed recognizable. In order to obtain a binary classification an optimal point was found for Task A by going about two thirds of the way towards D_{max} at 5.9 when employing Formula 3.4 for scoring the discourse enhanced concept sets. This means that all SRHs found below .59 on the inversely linear transformed score where subsequently classified as incorrect and those above as correct in this classification experiment. Given this setting and a word-concept ratio V of .39 the discourse contextually enhanced system yields the precision of $\approx .66$ on task T_A. This means that in 65.60% of all cases a markable - a speech recognition hypothesis - defined by the gold standard as correct is classified as such by the classification system. Which constitutes a gain of $\approx .14$ points over the corresponding baseline of $\approx .52$, which only classified 51.93% of the gold standard markables correctly. This classification result also lies .15 points below the human relative performance of .8.

Task B: Classification Performance

Again in this central task, as described in Section 3.2.2, the corresponding challenge is to determine the best speech recognition hypothesis from the set of

hypotheses that constitute more or less noisy representations of an original utterance. The classification system, consequently, will classify the highest scoring hypothesis with the value *best* and rest with *non-best* using the word to concept ratio of $V = .3$. The performance of the ensuing discourse context-sensitive classification system on this task was to $\approx .88$. That is, 88.07% of all cases where the best SRH defined by the human gold standard for T_B is among the best scored by the domain- and discourse sensitive algorithm. This constitutes a gain of $\approx .24$ points over the corresponding baseline of $\approx .64$, which only classified 63.91% of the gold standard markables correctly. This classification result still lies .07 points below the human relative performance of $\approx .95$.

Task C: Classification Performance

This task is described in Section 3.2.2 as well as in Section 3.2.2 and the corresponding classification challenge on the corpus SRH_1 is to differentiate coherent speech recognition hypotheses - which by themselves form a coherent utterance - from incoherent ones - where the no intention is deemed recognizable. In order to obtain a binary classification an optimal point was found for Task A by going about half of the way towards D_{max} at 4.4 when employing Formula 3.4 for scoring the discourse enhanced concept sets. This means that all SRHs found below .44 on the inversely linear transformed score where subsequently classified as incorrect and those above as correct in this classification experiment. Given this setting and a word-concept ratio V of .3 the discourse-enhanced system yields the precision of $\approx .71$. This means that in 71.05% of all cases a markable - a speech recognition hypothesis - defined by the gold standard as coherent is classified as such by the classification system. Which constitutes a gain of .08 points over the corresponding baseline of $\approx .63$, which only classified 63.05% of the gold standard markables correctly. This classification result also lies .09 points below the human relative performance of $\approx .8$.

Comparing the Performances

At this point, the presented classification experiments yielded two sets of performance data with their corresponding baselines, which allows for directly comparing the three task-specific performances of the domain context-specific system to those of the discourse-enhanced system. Given the six experiments performed on the task of hypothesis verification I will also present a statistical analysis thereof as well as looking at the potential contribution of this contextual computing approach to the robustness of the spoken dialog employed in the experiments. Initially, one can note that - despite the different concept sets that served as input in the discourse-sensitive run through the tasks - the results exhibit an internal consistency. In both runs he largest gains over the baselines are found in Task B, where including discourse context raises the gain over the baseline by .04 points, as the discourse enhanced algorithm classified 24.07% more of the markables correctly than the baseline whereas the solely domain-specific classified 20.06% more markables correctly. The second highest

Table 3.13: Domain and Discourse Context: Overview Experimental Results

classification task on corpus SRH_1	majority class baseline precision	domain algorithm precision	discourse algorithm precision	relative human precision
Accurate	≈ .52	≈ .65	≈ .66	≈ .80
BestOf	≈ .64	≈ .84	≈ .88	≈ .95
Coherent	≈ .63	≈ .70	≈ .71	≈ .80

gains are, again, to be found in both experiments on Task A, where including discourse context only raises the gain over the baseline by one point, but both the discourse enhanced algorithm as well as the solely domain-specific one classify respectively 13.16% and 13.67% more of the markables correctly than the baseline method. The smallest gain is found in Task C where discourse context again raises the gain over the baseline by one point and both algorithms classify respectively 7.3% and 8.04% more of the markables correctly than the informed majority-class baseline. The corresponding results are shown together with the human relative performances from the annotation experiments in Table 3.13.

The classification results presented in Table 3.13 also show that the discourse-specific enhancement yielded a greater improvement over the domain-specific system in the BestOf Task as compared to Tasks Accurate and Coherent. This, however, can be regarded as a result of the underlying task-specific questions. That is, in Tasks A and C the question of general correctness and coherence can be considered more domain- than discourse-dependent, while choosing between rival hypotheses for determining which one is best the inclusion of prior discourse context seems to provide valuable clues. While the data does not support a statistical analysis of this specific difference, it has, however, become feasible to analyze the overall gains of the experiments presented so far over their respective baselines. As the results of the second experimental run through the tasks are consistent - in the sense discussed above - with those of the first one, it has been shown that in all cases gains over the majority-class baseline can be achieved. In order to asses the overall likelihood that these six gains have arisen by chance it is possible to estimate their so-called *statistical significance*, which I will present in the following.

Analyzing the Performances

In order to analyze likelihood that the six gains presented in Table 3.13 are statistical significant, it is possible to view the two sample sets of performances - the baseline performances as sample A and the contextual computing performances as sample B, as a so-called *unpaired set* [Pearson, 1939]. The sets can be considered unpaired, as neither constitutes the classification system a further development of the baseline approach nor would one assume them to feature equal variances. Taking the opposite perspective, i.e. to view them as

Table 3.14: Statistical Analysis: Domain and Discourse Results

Sample	mean performance (95% confidence interval)	standard deviation from mean	median performance	average deviation from median
A baseline performances	59.6 (52.30 - 66.96)	5.99	63.0	4.02
B classification performances	74.0 (66.71 - 81.38)	9.7	70.7	7.01

task-specific pairs, would be statistically easier. Therefore, I will examine them as an unpaired set, given that the probability p of treating them as a pair will be lower than in an unpaired t-test.

Generally, the t-test - paired or not - has been designed to deal with small sample sizes by means of including a descriptive statistics report of the 95% confidence interval corresponding to the chance that is a *real* mean and standard deviation that one would find given a lager sample size, of which the observed mean and standard deviation may themselves be a deviation. Given the two samples of performances, presented in Table 3.13 a calculation of the corresponding unpaired t-test results in $t = -3.10$, given 10 degrees of freedom, which means that the probability of these gains is $p = 0.011$. This probability for assuming the null hypothesis is deemed statistically significant and based on the standard deviations, -errors and means as well as their confidence intervals displayed in Table 3.14. Also, as expected the resulting probability for a paired test is even more statistically significant and amounts to $p = 0.003$.

Comparison Shopping

Orthogonal to the statistical significance of the results gained by this contextual computing approach for scoring noisy speech recognition hypotheses, the question whether or not these gains suffice to improve the overall system performance and robustness can be examined by itself. In the following, I will examine the contribution of this inclusion of domain and discourse context in the light of existing scoring approaches, which are based on acoustic- and language models that - as described in Section 3.2.1 - are learned from previously recorded and transcribed data. As stated above and also shown in the discourse fragment consisting of the two sequential utterances given in Example 17 and Example 18, the central question in hypothesis verification from the point of view of overall system performance is to determine the best hypothesis to be selected for further processing. This, of course, corresponds to solving Task B.

(17) ich möchte auf dem schnellsten Weg von Köln nach Heidelberg

Table 3.15: Example hypothesis verification scores

hypothesis	statistical scores	domain and discourse scores
Example 19	1	.32
Example 20	.52	.48
Example 21	.34	.39
Example 22	.35	.00
Example 23	.52	.71

I want on the fastest way from Cologne to Heidelberg

(18) wie komme ich in Heidelberg weiter
how can I in Heidelberg continue

In the corpus of utterances and hypotheses SRH_1 the utterance, given in Example 18, yielded the hypotheses given in Examples 19 through 23.

(19) Rennen Lied Comedy Show Heidelberg weiter
Race song comedy show Heidelberg continue

(20) denn wie Comedy Heidelberg weiter
then how comedy Heidelberg continue

(21) denn wie Comedy Show weiter
then how comedy show continue

(22) denn wie Comedy weiter
then how comedy continue

(23) denn wie komme ich in Heidelberg weiter
then how can I in Heidelberg continue

The domain and discourse-based scores together with those of the automatic speech recognition system for the hypotheses - given in Examples 19 through 23 - are displayed in Table 3.15. The domain- and discourse score C'_{best} selects Example 23 as the contextually most plausible concept structure, i.e. more plausible than the alternative ones. This hypothesis was also labeled as the best SRH by the annotators.

Nevertheless, please note that a direct comparison of the performance of two computational approaches for $Task_B$ concerns the contribution of an explicit formal context model to the high performance of the automatic speech

Table 3.16: Performance Comparison for Task B

Classification System	Precision	Category
domain- and discourse model system	≈ .88	high-level context
domain model system	≈ .84	high-level context
acoustic- and language model system	≈ .84	low-level context
baseline system	≈ .64	majority-class

recognition system employed [Berton et al., 2006]. It is not to be seen as a proposal for a substitution of one with the other. In the terminology employed herein, all approaches that match the acoustic features extracted from the speech signal against acoustic and language models learned from training data [Schwartz and Chow, 1990] are viewed as low-level contextual computing techniques [Bunt, 2000], which can be employed variously to increase the robustness of natural language processing systems as displayed in Table 2.2. Their performance in this Task, as seen in Table 3.16, also exceeds the performance of the majority-class baseline approach derived from the gold standard by .20 points and is roughly equal that of the *high-level* system based on the domain model alone - the results are displayed in Table 3.13. Only the domain- and discourse-sensitive system exceeds it by classifying 4.19% more of the best hypothesis - as defined by the gold standard - correctly as such.

This first analysis of the results gained in this set of experiments on noisy speech data and their contribution within an advanced multimodal dialog system concludes the task of hypothesis verification, presented in Table 3.2. This task was employed as an example of *noise* generated by the speech recognition systems. The reason this type of noise lies in the multiple ways in which the features extracted from the speech signal can be mapped to lexically segmented phoneme sequences. As shown in the experiments presented above, domain and discourse knowledge can be employed to assist in resolving these phonetic ambiguities that arise in the *bottom-up* processing of human utterances. However, despite the significant results gained so far, several further intriguing research questions are raised by this approach of employing contextual knowledge.

Specifically, one can ask to what extend constitute the semantic paths - consisting of concept nodes and their semantic relations - appropriate representations of the meaning poles of the given forms. More generally, one can further seek to entangle the dependencies between the domain model given to algorithm and its performance. These questions concern both meaning construal and its formal representation, which fortuitously corresponds to the second and next challenge of those listed in Table 3.2, i.e. how to deal with the subsequent ambiguities in semantic interpretation. I will, therefore, discuss both the specific and general questions raised by the work presented so far in further experiments presented in the following Sections 3.4 through 3.6.

3.3.5 Roadmap

The six classification experiments presented in the last sections showed how contextual computing can be performed by recourse to ontological representations of domain knowledge and representations of entities form the discourse model. Specifically, they concerned the task of hypothesis verification, which was given as an example of noise due to alternative forms generated by speech processing. To further the overall examination of the possible contributions of higher level contextual computing as for discussing the additional questions facilitated through this approach, I will turn to the problem of ambiguity from the point of having presumably correct forms, but multiple mappings to possible meanings. Therefore, I will start with the well-known classification problem of word sense disambiguation - as exemplified in the lexical ambiguities for the form *bank* in Examples 1 through 2 - in the following Section 3.4.

3.4 Using Domain Context for Semantic Ambiguity

In the prior sections I have examined the specific gains of employing domain- and discourse context to *disambiguate* alternative form representations - in the form of speech recognition hypotheses. In this approach each hypothesis was mapped to multiple concept sets. Each concept set constitutes a different ontology-based representation of the potential discourse entities at hand [Byron, 2002]. Since these concepts correspond to names of nodes in a domain model - as described in Section 3 and Section 3.1.3 - they are connected via taxonomic *isa*-relations and non-taxonomic *semantic* relations.

The system employed for ranking these sets of concepts selects the best scoring concept set for a given hypothesis as the representative set for that hypothesis, thereby discarding the other sets as inferior mappings. Please note, that, for example, in Task B the hypothesis classified as best was the one with the highest scoring representative concept set. In an implicit manner the selection of a representative node set for the individual lexical forms contained in the hypothesis constitutes a semantic disambiguation of the form at hand if that form could have been mapped also to different nodes in the model and *semantic* is employed as it is in formal knowledge representation [Gruber, 1993], which I discussed in Section 3. This, of course, is pertinent as the algorithm employed for the domain- and discourse-specific ranking takes - in that sense - disambiguated - sets of concepts as input and in cases of semantic ambiguities ranks them depending on their average pairwise semantic path-length in the given domain model.[6]

I have discussed the importance of semantic ambiguity for scientific examinations of natural language in Section 2.3 as well as for natural language processing in Section 2.4. Until the advent of multi-domain spoken dialog system,

[6]Please note, that I will examine the importance of the semantic relations as well as the effects of the domain model in Sections 3.5 and 3.6 hereafter.

the problem of lexical ambiguities first and foremost constituted a challenge for text processing applications, e.g. information retrieval systems [Weiss, 1973], free text understanding [Sussna, 1993] as well as for evaluations of the corresponding disambiguating classification systems [Edmonds, 2002]. In the following, I will examine the selection of representative semantic representations as such a classification task, the main difference being that our markables are not typed, but correspond to the spoken language data as described in Section 3.2.2. The specific question of this examination of semantic ambiguities concerns the performance of the so far *implicit* selection of a correct mapping to a logical form, i.e. the appropriate node in the domain model.

The corresponding experimental setting employs an additional corpus of annotated speech data to be presented below as well as the domain model described in Section 3.1.1 to measure the reliability of the domain context-sensitive scoring approach in terms of its semantic disambiguation of the individual word senses found in the task-specific gold standard derived from the corpus. After localizing this approach to word sense disambiguation in the state of art, I will apply the evaluation methodology employed in Section 3.2 and describe the task-specific contribution of a corresponding classification system for word senses. Also, preluding the further examination of the semantic relations, an additional employment of the hierarchical nature of the semantic slots - discussed already in Section 3.1.3 will be presented along with the necessary algorithmic extensions. Thereby considering the semantic specificity of the non-taxonomic relations used for scoring the individual alternative sets of concepts that result from ambiguous word to concept mappings.

3.4.1 The Task: Word Sense Disambiguation

As in the case of the previous tasks it is again possible to employ both learned models derived from annotated data as well as knowledge-driven approaches to word sense disambiguation (WSD). In WSD approaches of the past can be divided into two types, i.e., data- and knowledge-based approaches. IN this case data-based approaches extract their information directly from textual corpora and are follow the common division into supervised and unsupervised methods [Stevenson, 2003]. While supervised methods work with a given set of potential classes in the learning process, e.g. stemming from a thesaurus [Yarowsky, 1992], or hand annotated data [Weiss, 1973]. As always, supervised methods require a manually annotated learning corpus. Unsupervised methods on the other do not determine the set of classes before the learning process, but through analysis of the given data by identifying clusters of similar cases, e.g. using clustering by committee [Pantel and Lin, 2003], which automatically discovers word senses from text given large amounts of data. In the case of spoken dialogue and speech recognition output the amounts of transcribed data are increasing, still more from research projects that record and transcribe speech data [Shriberg et al., 2004] than from commercially deployed spoken dialogue where access to the real data is restricted for legal reasons.

Still, given the basic distinction between data stemming from written cor-

pora[7] and data from spoken dialog systems, I categorized the latter further into controlled and conversational spoken dialogue systems in Section 1. Neither data- nor knowledge-driven word sense disambiguation have been performed on speech data stemming from human interactions with dialogue systems, since multi-domain conversational spoken dialogue systems for human computer interaction have not existed in the past. Now that some speech data from multi-domain systems have become available, experiments have been performed on that data ranging from anaphora resolution [Strube and Müller, 2003] via domain recognition [Rüggenmann and Gurevych, 2004a] to the annotation and classification experiments for the sense disambiguation discussed below.[8]

For this disambiguation task, i.e. Task D, the underlying question was to what extend the knowledge-based method for computing a domain-specific ranking described in Section 3.2.3 disambiguates correctly between alternative interpretations, i.e. concept representations, of a given recognized utterance at hand. For example, in speech recognition hypotheses containing forms of the German verb *kommen*, i.e. (to) come, the scoring approach described above obtains different sets of concepts - in one case the form *kommen* is mapped to the concept *MotionDirectedTransliteratedProcess*, which corresponds to the motion sense of the word or a *WatchPerceptualProcess* in the showing sense - given the vocabulary of the labeled nodes from the domain model.

Prior knowledge-based approaches have employed both lexical recourses as well as formal ontologies. The kind of knowledge therefore varies between more and less light-weight machine-readable domain models. In that respect the knowledge-based approach employed herein also has been tested with an ontology that is partially derived from lexicographic analysis found in the FrameNet data [Baker et al., 1998] as described in Section 3.1.3. In that respect comparable approaches employed WordNet as a lexical [Miller et al., 1990, Sussna, 1993, Voorhees, 1993]. Both employed taxonomic distances via hypernymy and synonymy and other relations, where modeled, between a number of input lexemes. Their disambiguation results on textual data also turned out to be significantly better than chance. Given the speech data described in Section 3.2.2 to be annotated with corresponding word senses taking from the ontological vocabulary I will possible to asses how the contextual computing system described in Section3.2.3 fares as a word sense classification system on a corpus of speech recognition hypothesis. Therefore, I will present the corresponding results based on the same method and metrics as in the corpus-based evaluations employed in the prior tasks.

[7] In that respect data-driven WSD was applied to various tasks, such as machine translation, information retrieval, content and grammatical analysis [Ide and Veronis, 1998] and other specifically designed collections of documents, as in the case of the SENSEVAL word sense disambiguation competition [Edmonds, 2002].

[8] It should be mentioned that prior WSD work on generation of spoken language exists with regards to finding correct phonetization of words in the field of speech synthesis where both supervised and unsupervised machine learning techniques were employed [Yarowsky, 1995].

Table 3.17: Domain Corpus - WSD_1

Corpus Name	WSD_1
Data Collection	Wizard of Oz Test
Subjects	224
User Utterances Transcribed/Annotated - turns	3100
Ambiguous Markables - SRH	2225

3.4.2 The Data: Collection & Annotation

Firstly, an annotation of the speech recognition hypotheses had to be performed to provide a gold-standard for evaluation as well as for the baseline computation. For this separate human annotators sense-tagged the data stemming from log files of the automatic speech recognition system, implemented in the SmartKom system [Wahlster et al., 2001], introduced as corpus SRH_1 in Section 3.2.2. To characterize this data from the point of view of sense disambiguation it is important to point out that there are notable differences between disambiguating spontaneous speech and texts, i.e., a smaller size of processable discourse context as well as hesitations, disfluencies and speech recognition errors.

Existing spoken language understanding systems can produce syntactic and semantic representations for multiple domains, e.g. the production system approach described by [Engel, 2002] or unification-based approaches described by [Crysmann et al., 2002], have shown to be more suitable for well-formed input but less robust in case of imperfect input. For conversational and reliable dialogue systems that achieve satisfactory scores in evaluation frameworks [Walker et al., 2000, Beringer et al., 2002] as described in Section 2.5, robust methods for disambiguating the sometimes less than ideal output of the large vocabulary spontaneous speech recognizers are, therefore, quite desirable.

In order to find a sufficient number of ambiguous markables new data set WSD_1 was taken from another corpus of SmartKom data that featured several additional domains, such as electronic program guide for TV and cinema information as well as assistance domains such as reservation or seating that were not given in corpus SRH_1. By means of the Wizard-of-Oz paradigm [Francony et al., 1992] a set of experiment were performed - where a full-blown multimodal dialogue system was simulated by a team of human hidden operators - with 224 subjects that produced 448 dialogues [Schiel et al., 2002]. After manual segmentation of the data into utterances corresponding to single intentions, e. g. a route or sight seeing request, the resulting audio files were transcribed. Then, the segmented audio files were again given to the speech recognition engine integrated in the dialogue system [Wahlster, 2003]. The corresponding speech recognition word lattices [Oerder and Ney, 1993] were transformed into n-best lists of speech recognition hypotheses. For obtaining the data set WSD_1 a random sample of 3100 utterances was taken for the annotation experiment as displayed in Table 3.17, which contained 2225 ambiguous markables. Again,

Table 3.18: Task T_D Disambiguation - Annotation Experiment - WSD_1

Data Source Corpus WSD_1		
Task	Classification Values	Human Performance
T_D	word senses	precision $\approx$ .79

.

the utterances containing no ambiguous forms could have been removed from the corpus, however, since all performance measures in this case are based not on the utterances as markables, but on the ambiguous forms, they do not affect them. Additionally, the corpus will, again, be employed further.

The annotation of the data was done by two persons specially trained for the annotation tasks, again with the purposes of, firstly, assessing relative human performance in terms of measuring inter-annotator reliability. Secondly, another gold-standard is needed for this task to evaluate the classification systems' performance. For that purpose, the annotators reached an agreement on annotated items of the test data on which they had differed in the first place. The resulting gold-standard, therefore, represents the highest degree of correctly disambiguated data employable for comparison with the tagged data produced by the disambiguation system.

Task D: Annotator Performance

As in the case of Task B a class-based kappa statistic cannot be applied here, as the classes vary depending on the number of mapping per ambiguous form to the ontology. For the annotation task, corresponding forms where automatically generated and displayed to the annotators [Müller, 2002]. Also an additional class, i.e., `not-decidable` was allowed for cases where it is impossible to assign sensible meanings. The WSD_1 data set altogether was annotated with 2225 markables of ambiguous tokens, stemming from 70 ambiguous words occurring in the test corpus. Concerning the question whether humans are able to annotate the data reliably or not, the former is the still case despite the problematic nature of the data, as shown by the resulting inter annotator agreement of 78.89%. This measure can also be regarded as the relative human performance introduced in Section 3.2.2 and is shown in Table 3.18.

3.4.3 The Algorithm: Scoring Word-Sense Ambiguities

As before, the classification system performs a number of processing steps as described in Section 3.2.3: the first preprocessing step is to convert each SRH into a concept set (C). For that purpose the system's lexicon can be used, which contains either zero, one or many corresponding concepts for each entry. A simple vector of concepts - corresponding to the words in the SRH for which entries in the lexicon exist - constitutes each resulting set. All other words with empty concept mappings, e.g. articles, are ignored in the conversion. Due to

Table 3.19: Example mappings of forms to concept labels

Linguistic Forms	Labels of Concepts
Ich	Person
bin	StaticSpatialProcess SelfIdentificationProcess NONE
auf	TwoPointRelation
Philosophenweg	Location

Table 3.20: Example alternative concept sets

C_1	{Person, StaticSpatialProcess, Location}
C_2	{Person, StaticSpatialProcess, TwoPointRelation, Location}
C_3	{Person, SelfIdentificationProcess, Location}
C_4	{Person, SelfIdentificationProcess, TwoPointRelation, Location}
C_5	{Person, TwoPointRelation, Location}
C_6	{Person, Location}

lexical ambiguity under scrutiny herein, i.e. the one to many word - concept mappings, this processing step yields a set $I = \{C_1, C_2, \ldots, C_n\}$ of possible interpretations for each hypothesis.

(24) Ich bin auf dem Philosphenweg
I am on the Philosopher's Walk

For example, the words occurring in example utterance given in Example 24 feature the word-specific mappings to the nodes of the domain model that are shown in Table 3.19. Herein all forms with a single mapping - in this example they are *Ich* (I), *auf* (on) and *Philosophenweg* (Philosopher's Walk) - are not considered as markables. Only forms with multiple mapping such as *bin* constitute markables to be tagged with the respective concept labels as the attribute *word sense*. In this case *bin* is mapped either to *StaticSpatialProcess* - which is the appropriate sense - it is also mapped to *SelfIdentificationProcess* and *None* meaning it is understood as a self referential as in *Ich bin Robert* (I am Robert) or as a grammatical marker, e.g. marking perfected processes in German.

The task of the scoring algorithm presented in Section 3.2.3 is to assign a domain context-specific value to each possible interpretation in the set I. If the highest scoring concept set C_s contains the correct sense of the ambiguous form as defined by the gold standard, then a correct classification of the word sense has been performed. Due to the multiple concept mappings for the form *bin* the resulting set of concept representations I for Example 24 is shown in Table 3.20.

Please note that the individual concept representations can consist of a different number of concepts, as the mapping to *None* is not represented in the individual conceptual representations. As mentioned above, the mapping *None* is assigned to lexemes which constitute potential functional grammatical markers, however, this is not to say that there no other potential senses, than the mappings specified for the SmartKom domains. As a matter of fact, numerous other senses - stemming from other domains - could be added and - due to language use and change - new mappings can arise dynamically and become more or less entrenched in a community of speakers [Langacker, 2000] as well as in communities of language learning agents [Steels, 1998a, Steels, 1998b]. Nevertheless, since the collected language data that produced the corpus WSD_1, contains no out-of-domain utterances or require novel construals in the in-domain ones from the point of view of the domain model used to classify the meanings that occur in the data, it would hardly make sense to include them or allow for their dynamic inclusion at this point.

As in the previous experiments, the classification system converts the domain model, i.e. the ontology described in Section 3.1.3, into a directed graph with concepts as nodes and relations as edges. In order to find the shortest path between two concepts and score their semantic connectivity the algorithm, shown in Section 3.2.3, was employed and also fitted with a new addition. Please note again, that the ontology employed for the evaluation bases on a generic top-level ontology and a modeling of *Processes* and *Physical Objects* as a kind of event that is continuous and homogeneous in nature, The semantic relations base on the frame semantic analysis used for generating the FrameNet data [Baker et al., 1998]. The hierarchy of *Processes* is connected to the hierarchy of *Physical Objects* via slot-constraint definitions herein referred to as *semantic relations*.

Given the importance of semantic relations for calculating their degree of semantic connection in this approach and the fact that these relations are modeled in a so-called *slot-hierarchy* themselves, it is possible to assign different weights to the individual relations found by the algorithm, depending on their level of granularity within the relation hierarchy. That means, for the top relation *has-role* a weight of 0 is assigned, for all direct sub-relations of that relation a weight of -1 is assigned, consequently the weight is decremented by 1 for each further decent down the relation- or slot-hierarchy, i.e. a weight of -2 the sub-sub-relations of the top one, -3 for the sub-sub-sub-relations and so forth until the lowest branch of the tree has been reached. For example, a broad level relation such as *has-theme*, as found in the class statement of *Process*, is weighted with -1 as it has only one super-relation, i.e. *has-role*, whereas a more specific semantic relation, such as *has-actor*, is weighted with -4 because it has four super-relations, i.e. *has-artist*, *has-associated-person(s)*, *has-attribute* and *has-role*, in the hierarchy.

As before, the algorithm selects from the set of all paths between two concepts the one with the smallest weight, i.e. the *cheapest*. The distances between all concept pairs in CR are summed up to a total score. Note, that more specific relations subtract more then less specific ones from the aggregate score. The

set of concepts with the lowest aggregate score represents the combination with the highest and most specific semantic connectivity with respect to the domain model. Following an assessment of the baseline performance for this task, I will present the classification results gained by employing both the original scoring approach employed so far as well as the approach including the described weights for the semantic relations found between the concepts.

3.4.4 The Results: Word Sense Disambiguation

For this evaluation of algorithm the dataset of corpus WSD_1 - presented in Table 3.17 was employed. Based on the double human annotations - presented in Table 3.18 - for the given classification task, a corresponding *gold standard* was crafted via inter-annotator negotiations. Given this gold standard mark-up for the Tasks Disambiguation (T_D), a corpus-based computation of a baseline metric has again become feasible, as I will describe below.

Task D: Baseline Performance

In order to stay consistent with the methodological framework employed herein, a *proportional* majority class baselines can be computed as described in Section 3.2.4 . Hereby, all markables in the gold-standards were counted for computing the total number of values for a markable as defined as in Formula 3.6. Based on this the most frequent attribute for that markable is computed as shown Formula 3.8. Finally, it is possible to compute the task-specific proportional baseline as defined in Formula 3.9 This approach, then, yields the percentage of correctly chosen concepts by means of selecting the most frequent meaning without the help of a system as described in Section3.2.4. This approach manages to tag 52.48% of the markable correctly, as defined by the gold standard, resulting in a baseline performance of $\approx .52$ for corpus WSD_1.

Task D: Classification Performance

The percentage of correctly disambiguated lexemes from both systems is calculated as given in Formula 3.10.

$$R = \frac{g + n}{w * 100} \tag{3.10}$$

Where R is he result in percent, g the number of lexemes that match with the gold-standard, n the number of not-decidable ones and w the number of total lexemes. As both systems never score *not-decidable*, any chosen concept is scored positively for these cases equally for all approaches. Again, for this evaluation a score was computed for each concept set in I using Formula 3.4 with the, by now, standard setting of discarding hypotheses whose word-to-concept ratio is above 3. The concepts in the highest ranked set are considered to be the ones classified as the correct word sense in this context by the system.

Keeping in mind that in this approach the relations between two concepts are weighted D_{max} for solely taxonomic relations among concepts and 1 for each

Table 3.21: Results Word-Sense Disambiguation - Classification Experiment

classification task on corpus WSD_1	majority class baseline precision	$domain_O$ algorithm precision	$domain_V$ algorithm precision	relative human precision
Disambiguation	$\approx .52$	$\approx .64$	$\approx .65$	$\approx .79$

semantic relation connecting them. The alternative approach described above assigns each relation an individual weight according to their level of generalization. Compared to the gold-standard of task T_D, the first approach already classified 63.76% of the markables correctly, yielding a precision of $\approx .64$. Additionally a similar gain over the baseline can be reported for the alternative approach, which classified 64.75% of the markables correctly, yielding a precision of $\approx .65$. While the alternative approach brought only a slight gain by interpreting about 20 lexemes more - out of the 2225 ambiguous ones contained in the data - than the original version, both the original $domain_O$ and the variant $domain_V$ systems managed to exceed the baseline performance of $\approx .52$ by .12 and .13 points respectively, as shown in Table 3.21. Again, these classification results fall .15 and .14 points short of the relative human performance of $\approx .79$.

Analyzing the Performances

Looking at these results as an additional examination of the contribution of contextual domain and discourse knowledge as compared to an informed majority class baseline, it is possible to add these classification and baseline results to the respective samples A and B for which a statistical significance test has been performed and described in Section 3.3.4. Given these increased samples of performances, i.e. those presented in Table 3.13 and Table 3.21, a calculation of the corresponding unpaired t-test results in $t = -3.52$, given now 14 degrees of freedom, which means that the probability of these gains drops even below that of $p = 0.011$ reported for the hypotheses verification gains alone to $p = 0.003$.

At the end of Section 3.3.4 I posed several question that constituted the onset of this examination. The first concerned the amount of correct word to concept mappings found in the best scoring concept set, which - as described above - can be considered significantly above the majority class baseline performance seen together with the prior already significant gains of this approach. Having examined this first questions, the next question in line concerns the amount of correct semantic relations found in the network connecting the individual concept pairs of the best scoring concept set. This second specific question will be examined in the next section of this chapter, i.e. Section 3.5 below, followed in Section 3.6 by an analysis of the more general question concerning the role of the domain model that serves as a representation of that type of context for the experiments presented herein as well as *vice versa*.

3.5 Using Domain Context for Relation Extraction

The findings presented in Sections 3.2, 3.3 and 3.4 based on an approach to employ domain- and discourse context to rank alternative form representations of a speaker's utterance - presented to the system as multiple speech recognition hypotheses of that utterance. Hereby, each hypothesis was mapped to multiple concept sets as a result of lexical ambiguity. As described above, the system employed for ranking of sets of concepts bases on the average path length between all concept pairs as defined in Formula 3.4, whereby the best scoring concept set for a given hypothesis is taken as the representative set for that hypothesis. As pointed out this implicit selection of a representative node set for the individual lexical forms contained in the hypothesis can be regarded as performing a semantic disambiguation task of the ambiguous forms at hand. Concerning the corresponding question of the system's performance on this task, an examination thereof has been presented above in Section 3.4 yielding results that reinforce the significance of the gains achieved in the experiments on noisy speech hypothesis when put together in an unpaired t-test.

Nevertheless, I have also noted that, next to the specification of an appropriate concept mapping, the semantic relations that hold between the individual concept pairs are extracted from the ontology thereby creating the noted *semantic paths* that consist of nodes connected via the extracted semantic relations. This, in turn, poses the corresponding performance question when assuming this as a relation extraction task for the system, given that a corresponding disambiguation - as discussed above have taken place. Additionally, looking at these specific questions concerning the adequacy of the node concepts and the relational arcs between them a more general question concerns the dependencies between the domain model and the algorithm and its performance results. Essentially, this general question can be paraphrased by asking what if the domain model employed would have featured other conceptual class divisions and corresponding mappings to lexical forms or if other semantic relations would have been modeled. In other words one can even ask what it say about the domain model, i.e. the formal knowledge representation [Gruber, 1993] of the contextually evoked domain, that some incorrect concept mappings and - as I will discuss below - some incorrect relations are found in the best scoring semantic paths.

In the following, I will, therefore, first present an examination of relation extraction and the corresponding system performance in addition to the disambiguation and the noise-related classification tasks discussed above. Then, concluding this analysis of domain and discourse context, I will examine the question how these results reflect back onto the given model of domain and discourse knowledge. Furthermore, as I will discuss below, this reflection of the *faults* of the model - seen in the *mirror* of the obtained performance results - in turn raises the more general question if such task-specific evaluations could also be employed to the benefit of the given representation of the context at hand. While the first question can be approached on the basis of the methodology

employed before, the latter one requires new methodological and analytic considerations, which I will present after discussing the remaining task of relation extraction, as Task E in the following section.

3.5.1 The Task: Relation Extraction

The task of extracting the semantic relations from a domain model that are then regarded to hold between the entities denoted by the given forms, is comparable to work on *role-labeling* [Gildea and Jurafsky, 2002, Màrquez et al., 2008]. Moreover, in the case of the initial work performed by Gildea and Jurafsky (2002), the set of role labels used corresponded to so-called *frame roles* found in the annotated FrameNet corpus [Baker et al., 1998], which, as described in Section 3.1.3 also served as a *role model* for the relation hierarchy employed herein. Role-labeling has meanwhile also been employed to support other natural language processing tasks, ranging from coreference resolution to question answering [Ponzetto and Strube, 2006, Shen and Lapata, 2007]. Additionally the task discussed herein features similarities to the scenario template task of the Message Understanding Conferences [Marsh and Perzanowski, 1999]. In this case predefined templates are given, e.g. *is-bought-by(COMPANY A,COMPANY B)*, which have to instantiated correctly, e.g. in the phrase such as given in Example 25 the specific roles, i.e. Polygram as COMPANY B and Island Records as COMPANY A have to be put in their adequate places within the overall template.

(25) Polygram has bought Island Records (BNC:A1E 465)

Relation extraction, as such, primarily refers to corpus- and pattern-based approaches for extracting semantic- [Hearst, 1992, Cimiano et al., 2005] or taxonomic relations [Rosenfeld and Feldman, 2006, Blohm et al., 2007] from natural language texts for, by and large, semi-automatic ontology learning. Again, context in these approaches bases on co-occurrences of lexical items and syntactic parse trees [Gildea and Jurafsky, 2002, Zhou et al., 2007] available for textual data. Given the semantically annotated data described above, stemming from multimodal interaction with a *Wizard-of-Oz*-based conversational multi-domain dialogue systems, it is possible to perform parts of the empirical evaluation experiments as undertaken above. This examination will, again, include a corresponding annotation- and classification experiment as a further performance evaluation of the contextual computing approach as a relation extraction system.

3.5.2 The Data: Collection & Annotation

For this annotation task only a tenfold of the WSD_1 data set was employed. These 10% were taken from the hypotheses that had been identified as being the best - given the approach described in Section 3.2.1. This provided 977 non-taxonomic semantic relations posited between the concepts contained in the respective sets as markables extracted by the system. Please note, that - as in

Table 3.22: Domain Corpus - REL_1

Corpus Name	REL_1
Data Collection	Wizard of Oz Test
Best Speech Hypotheses Annotated -	200
Initial markables - Relations	977

the case of Example 26 below - there can be a *chain* of relations connecting two given concepts, the ramification of which I will discuss below in Section 3.5.3. An overview of the resulting corpus, REL_1, and the initial markable set is given in Table 3.22.

For these utterance-based representations of the semantic relations that predicate the concepts that are part of the ontology's event hierarchy the question arises concerning an appropriate annotation scheme for labeling the semantic relations. As I will discuss below, *incorrect* relations could be ones that are extracted instead of the correct one, ones that are missing and ones that are superfluous. Also, let me note the first, but not the last, of the differences in this examination, which stems from the fact that the annotators were given the relations extracted by the system as markables, instead of annotating *raw* data with the same values as the system will. As this also has a methodological bearing for measuring the corresponding system performance, I will introduce the annotation scheme devised along with its consequences for the evaluation.

Methodological Metric

For evaluating the performance of the system I defined three types of values for labeling incorrectly extracted relations and one for correct ones. The four resulting values to be attributed to the markables were:

- deletions, i.e. missing relations in places were one ought to have been identified;
- insertions, i.e. postulating any relation to hold where none ought to have been;
- substitutions, i.e. postulating a specific relation to hold where some other ought to have been;
- identifications, i.e. if the correct semantic relation (role) was extracted by the system for the corresponding concept pair.

The first three values, deletions, insertions and substitution constitute a transfer of the word error rates employed in speech recognition. Since they were described in Section 2.5 I will proceed to exemplify their application in this domain.

Table 3.23: Task T_E Extraction - Annotation Experiment - REL_1

Data Source Corpus REL_1		
Task	Classification Values	Human Performance
T_E	{substitution,deletion,insertion}-identification	precision ≈ .80

.

An example of a substitution error in this task-specific annotation scheme is given the hypothesis shown in Example 26.

(26) wie komme ich von hier zum Schloss
how come I from here to castle

In this case the sense disambiguation - described in Section 3.4 was accurate, so that the two ambiguous entities, i.e. *kommen* (to come) and *Schloss* (castle), were correctly mapped onto a *MotionDirectedTransliterated*-process and a *Sight*-object - the remaining concept Person resulted from an unambiguous word-to-concept mapping for the form *ich* (I). The error in this case was the substitution of the appropriate *has-goal* relation with the extracted relation *has-source*, as depicted in Figure 3.6.

As a special case of substitution the annotators were instructed to count those cases as inaccurate where a relation chain was selected by the algorithm. While in principle such chains, e.g. metonymic chains are possible and in some domains not infrequent, in the still relatively simple and short dialogues that constitute corpus REL_1 they do not exist. Therefore cases, such as the semantic path between *WatchPerceptualProcess* and *Sight* shown in Example 27 were also counted as substitutions, because simpler ones should have been extracted or modeled, as shown in Figure 3.7.

(27) ich will das Schloss anschauen
I want to see the castle

The annotators were instructed to mark deletion errors in such cases where a gold standard annotation of all concept pairs should have extracted a relation, e.g. if no semantic path and *WatchPerceptualProcess* and *Sight* in the case of the Example 27 would have been extracted as depicted in Figure 3.8 . This mark-up requires the same understanding of the domain-specific meaning of the utterances as in the opposite case, where annotators had to mark insertion errors, i.e. where any semantic path, e.g. between [Agent] and [Sight] in Example 27, were extracted by the system as shown in Figure 3.9. The inter-annotator agreement on this task amounted to 79.54%, shown as the relative human performance in Table 3.23 given the values substitutions, deletions, insertions and identifications.

Table 3.24: Results Relation Extraction Experiment

extraction task on corpus REL_1	majority class baseline precision	domain algorithm precision	relative human precision
Extraction	not applicable	≈ .76	≈ .80

3.5.3 The Results: Relation Extraction

The data for this evaluation was produced by the system described in Section 3.4.3 from which a subset was turned into a gold standard for this experiment, using the annotation scheme described above. Other than in the previous settings the creation of a gold standard through *corrective* annotation using this scheme slightly changed the number of markables. This is the case as each deletion error calls forth an - as of yet unspecified - semantic path as a markable that could be a correct identification, which was left blank by the system. On the other hand semantic chains that were conflated in the annotation into a singular relation, as in the case of the substitution error given in Example 27, reduces the number of markables. As a result the gold standard corpus contains four markables less than given in Table 3.22 displaying the initial count of semantic relations extracted by the system, due to the fact the there were four more conflations than deletions, as I will show below. The performance data, therefore, is based on 973 markables, representing - as before - the ideal solution.

As compared to this gold standard 76.31% of the relations extracted by the system correctly identified the semantic relation between the concepts - using the system and settings described in Section 3.4.4. In 23.69% of the cases one of the three extraction errors was found by the annotators. I will discuss and examine the distribution of these specific error types in Section 3.6. For concluding this specific examination of the relation extraction performance, another difference in this evaluation as compared to the previous ones needs to be noted first. This difference concerns the calculation of a corresponding baseline. In the case of $Task_E$ a computation of a corresponding majority class baseline has been thwarted, as this baseline approach requires to compute how many markable tokens assume the most frequently given value as compared to the rest of the tokens. In this case, the annotators were deliberately not asked to determine alternative semantic relations for the incorrectly extracted ones, which would have involved delving deeper into the given ontology and its engineering principles, than feasible for an annotation task. Given the difficulties in calculating the needed token-based frequencies and determining the markable-specific values, in terms of specific alternative relations and their distribution, this performance result of a precision of ≈ .76, as shown in Table 3.24 cannot be included in the statistical analysis of the respective gains over majority class baselines presented in Sections 3.3.4 and 3.4.4.

Performance Reflection

However, both the performance of the system in this task as well as that of the annotator be seen especially encouraging in several respects. For one it shows, that over three quarters of the semantics paths - on which the algorithm described in Section 3.2.3 substantially bases its computation as shown in Formula 3.4 - are reliably regarded to identify the correct semantic relation between the given concepts in that context. While direct comparisons with the prior experiments might be misleading due to the differences inherent this experimental setting, a more intriguing perspective arises in the analysis of the remaining quarter, which - as stated above - is motivated by the corresponding question how much these specific errors reflect shortcomings of the model used in the evaluations rather than in the algorithm for scoring the sub-graphs extracted from that model. In the following Section 3.6, I will, therefore, discuss this final question regarding the explicit formal model used to represent domain and discourse context below in Section 3.6 before concluding this chapter.

3.6 Evaluating Domain Context

The reliance on an existing ontological model of domain knowledge - as pointed out above - on which the findings presented in Sections 3.2, 3.3, 3.4 and 3.5 hinge, raises the question of how to evaluate such representations of domain context. I, now, address this question providing new methodological and analytic considerations regarding the possibility to evaluate the quality of a given explicit context model.

The need for the establishment of evaluation methods that can measure respective improvements or degradations of ontological models, e.g. yielded by a precursory ontology engineering stage - be it manual development or automatic learning - is undisputed. I will, therefore, present an methodological framework that - in principle - allows to evaluate a number of different domain models in terms of their *performance* on specific tasks.

The resulting task-based approach for quantitative ontology evaluation also opens the door for a bootstrapping approach to ontology engineering. This approach relies on the fact that tasks commonly feature a so-called *gold-standard* defining perfect performance, as discussed above. By selecting ontology-based approaches for the respective tasks, the ontology-dependent part of the performance can, theoretically, be measured. Following a general presentation of this approach, I will how the results of the prior experiment, discussed above in Section 3.5.3, can be re-cast as an evaluating of the context model used.

Moreover, as the employment of ontologies has gained in importance for the development of intelligent systems, services and applications, questions concerning their evaluation also moved correspondingly into the foreground. In the area of natural language processing alone, for example, ontologies have successfully been used to represent pertinent domains of interest and to provide knowledge for variety of tasks as described in Section 3.1, including - now -

those described in Sections 3.2 through 3.5. Broadening the perspective, such formal model have also been used for authoring natural language processing systems, e.g. for defining interface specifications of multi-domain dialogue systems [Gurevych et al., 2003a].

Even higher up, ontologies constitute an integral part of the Semantic Web [Berners-Lee et al., 2001] and numerous other projects. I have discussed their formal properties and benefits in Section 3.1.1. Yet, many well-known problems remain. Two critical issues concern:

- the knowledge-acquisition bottleneck, where ontology learning and population come into play [Buitelaar and Magnini, 2005],
- the lack of formal means for evaluating the fitness of a given ontology or an ontology improvement in light of the task at hand [Gangemi et al., 2005].

The next and last task, therefore concerns the contribution of the work presented above on assessing the *fitness* of the underlying domain model for a given task.

3.6.1 The Task: Evaluating Ontological Fitness

For this task, I will follow the general distinction between qualitative and quantitative ontology evaluation [Brewster et al., 2004] and between descriptive and revisionary ontologies [Gangemi et al., 2001]. In the proposed framework to perform quantitative evaluations of descriptive domain models[9], which can also serve in ontology engineering as a form of incremental ontology improvement.

The underlying question in the proposed quantitative evaluation can be expressed by asking how fit a given ontology is for a well-defined task. In the same vein, any ontology change, which transforms an ontology from a state 0_n into O_{n+1}, can only be considered *successful* if the resulting ontology O_{n+1} is fitter than O_n in some tangible respect, which - as argued herein - can be an ontology-dependent task performance. Herein *fitness* is, therefore, employed in a straight-forward sense: if an ontology is to be used for a given task, e.g. scoring based on domain context as described in Section 3.2.3 - it can be used to perform better or worse, being fitter or less fit for it in a measurable way.

In order to measure different degrees of fitness, possible in case where the performance depends on the ontological model, the algorithmic side of the equation ought to be constant throughout an evaluation. A specific evaluation suite should therefore, be selected such that the measurable output concerning the given task depends as much as possible on the ontology used. It is important to point out that the type of ontology evaluation proposed herein can be carried out only with respect to a given task at hand, which the specific ontology has to *solve*. A task-independent automatic evaluation still remains an elusive goal

[9] In principle, the framework introduced herein should also be applicable to revisionary ontologies, but the experiments as well as the corresponding foci have rested on descriptive ontologies so far.

for which a general solution awaits discovery [Guarino and Welty, 2002].[10] The central methodological focus of this evaluation, therefore, lies on examining the feasibility to test and incrementally augment - i.e., successfully increase their fitness - ontologies given a well-defined problem based on a evaluation *gold-standard*. It is also feasible to elevate the test to a higher level of generality, by crafting *benchmark tasks* that are representative of a specific classes of problems.

The general evaluation framework introduced herein can encompass all of the ontological levels introduced below. Specific evaluations, however, could also examine an ontology on one or more of the following three basic levels independently:

- the fit of the vocabulary, i.e. usefulness of the ontology classes or *concepts*;
- the fit of the taxonomy, i.e. the usefulness of the *isa* hierarchy;
- the fit of the non-taxonomic relations, i.e. the usefulness of the *semantic* relations.

In principle, aggregate evaluations combining the various levels are also possible. More importantly, however, for this work is the proposal that meaningful transfers of the commonly used *error rates* - i.e. insertions, deletions and substitutions - exist for the domain of evaluating and populating ontologies, as discussed above in Section 3.5.1. As described in Section 2.5, these error rates are commonly used in automatic speech recognition [Jurafsky and Martin, 1991] and have been proposed as well as for evaluating the performance of concept- and relation taggers [Higashinaka et al., 2002, Gildea and Jurafsky, 2002].

The need to develop a clear set of evaluation methodologies is also widely acknowledged [Guarino, 1998], whereby the qualitative type of evaluations basically relies on user or expert judgments [Gomez-Perez, 1999]. Hereby it is left open whether ontology engineers, system users or domain experts ought to be the judges. Additionally, to judge ontologies in terms of the principles on which their design has been based also bases on criteria defined by the *external semantics* which, again, has to be evaluated by human experts. There are even more general problems have been discussed that arise from such principle-based approaches [Wilks, 2002].

Since the concern of this section is on quantitative evaluation for measuring the performance of an ontology for a given task, I will not discuss the valuable work on measuring similarities between ontologies [Hovy, 2001] or evaluating a given ontology against a pre-defined ontological *gold-model* [Maedche and Staab, 2002]. However, as I will discuss below, the potential ontology improvement that - in a sense - falls out of this evaluation is comparable to prior approaches to ontology learning and population with respect to the basic levels proposed above [Stevenson, 2002a], since it also enables an evaluation

[10] Task-independent evaluations might even be impossible in principle, as it is widely acknowledged that ontology engineering and employment has many task-dependent features and constraints, which can be considered as taking an interactional point of view in the sense of Dourish (2001).

of the fitness of ontological models for each level respectively and is independent from the automatic or manual means by which the ontology was crafted. The work presented herein is, moreover, in principle related and in perfect agreement to that of Brewster et al. (2004), who state that:

> The establishment of a clear set of simple application suites which would allow a number of different ontologies to be slotted in, in order to evaluate the ontologies would be an important research step. [Brewster et al., 2004]:2

In this work, they provide a data-driven approach that enables an ontology evaluation against given textual corpora. Employing this promising approach they also arrive at a measure for *ontological fit*. This constitutes a measure of vocabulary overlap between the concepts contained in a given ontology and the terms extracted, by means of a latent semantics-based clustering algorithm [Foltz et al., 1998], and expanded by means of a two step hyponym WordNet look-up. The necessary alignment or mapping between concepts and terms is performed by manual annotation[11]. For measuring the taxonomic fit the authors employ WordNet distances [Stevenson, 2002b], thereby, hitting on the so-called *tennis problem* [Hayes, 1999], that refers to the finding that some terms in WordNet are further apart than expected due to their dispersal by type.

On top of that, an ontology provides more than a vocabulary of entities and their generalization hierarchy. A substantial amount of its expressive and inferential capabilities (at least for natural language processing applications - as shown above - lies in the non-taxonomic relations that hold between the concepts. For evaluating this aspect of an ontological model no solution has been proposed so far. I will, therefore, examine the feasibility to fill this gap by the proposed performance- or task-based evaluations that can yield measures of how fit the vocabulary, taxonomy and the non-taxonomic relations are for a given task at hand.

Methodological Framework

In the following, I will sketch out the necessary elements that are needed for Task_F examining the fitness of given domain models. That is, to define the scope of the experiments, a metric for evaluating ontologies and a list of ingredients for performing corresponding ontology evaluations. As shown in Table 3.25, the classical scope of ontology learning and population approaches can be regarded as constructive learning, wherein new concepts and relations are learned. Employing the error rates introduced herein, this surmounts to reducing the amount of deletion errors. In prior work an additional distinction is made between ontology learning and population in case of learning instances. In both cases iterative additions are made - by means of some learning approach - to an initial ontological state O_n and a resultative state O_{n+1}.

[11] Unfortunately, the authors do not provide a measure for inter-annotator agreement on this task, which, as the data presented in Section 3.4.2 show is also not a trivial task.

Table 3.25: Scope of ontology learning (*X* denotes coverage and *O* the opposite)

levels - errors	insertion errors	deletion errors	substitution errors
vocabulary	O	X	O
isa relations	O	X	O
semantic relations	O	X	O

More importantly, however, I propose regard ontology improvement to include not only forms of constructive learning, but also - what could be considered - *destructive* learning for removing superfluously inserted entities and corrective learning operations for substitutions as well. Correspondingly, the scope of ontology learning and population is not limit to that of inserting new entries, but also can include acts of deleting existing parts and performing corrective substitution operations. An appropriate general term that covers additive, subtractive and other substitutive operations might be ontology *crafting*. Regardless of naming conventions an expressive metric is needed for evaluating ontologies or the potential improvements brought about by specific crafting operations.

Evaluation Metric

In the past different learning and population approaches were applied respectively for the three basic levels of vocabulary (level 1), taxonomy (level 2) and (non-taxonomic) semantic relations (level 3), as shown in Table 3.26 . In much the same way these levels have been subject to independent evaluation approaches (for an overview see also Table 3.26). For the evaluation and population framework described herein, I propose that the notion of error rates - common in evaluations of automatic speech recognition performance as word error rates [Jurafsky and Martin, 1991], but also known from previous work on concept tagging as concept error rates [Higashinaka et al., 2003] and discussed in Section 2.5.2 and applied in Section 3.5 - can, furthermore, be transferred for evaluating each of the ontological levels displayed in Table 3.27.

Therefore, the results of a task-based evaluation should display the following shortcomings:

- *insertion errors* indicating superfluous concepts, isa- and semantic relations;
- *deletion errors* indicating missing concepts, isa- and semantic relations;
- *substitution errors* indicating *off-target* or ambiguous concepts, isa- and semantic relations.

Given appropriate tasks and maximally independent algorithms operating on the ontology in solving these tasks in conjunction with task-specific evaluation

Table 3.26: Approaches to OLP and Ontology Evaluation approaches

Level L_1	
concepts vocabulary	learning [Pereira et al., 1993, Stevenson, 2002a] evaluation [Brewster et al., 2004, Gomez-Perez, 1999]
Level L_2	
hierarchy granularity	learning [Widdows, 2003b] evaluation [Stevenson, 2002a]
Level L_3	
semantic relations	learning [Gildea and Jurafsky, 2002, Ciaramita et al., 2005] evaluation see Section 3.6.3 and [Porzel and Malaka, 2004a]

gold-standards, one can calculate the error rates corresponding to specific ontological shortcomings. The general semantics of the level-specific error types are given in the overview of the proposed transfer of these error rates to the three basic ontological levels displayed in Table 3.27.

With this, I can provide performance measures that can:

- evaluate one or more ontologies in terms of their *performance* on a given task (ideally to measure only the ontology-specific aspect of the performance),
- quantify the respective gains and losses of the insertion, deletion and substitution errors,
- populate (re-craft) the ontology as derived from the individual error type specific results, and
- re-evaluate the respective performance in- or decreases resulting from the crafting operations.

By applying this evaluation scheme one can, therefore, test and measure the respective improvements that are brought about by individual learning and population approaches that target the individual levels. Furthermore, one can also categorize and compare these approaches as shown in Table 3.26.

Table 3.27: Task: Ontology Evaluation: Levels & Error Types

level	insertion	deletion	substitution
1	irreverent concepts	omitted concepts	ambiguous concepts
2	isa too coarse	isa too fine	isa too polygamous
3	irreverent relations	missing relations	indirect relations

The Evaluation Ingredients

Next, I will specify the minimal elements and their specific constraints that are necessary for a task-based evaluation of an ontology and its entire range of relations. An overview of such a generic task-based evaluation suite is given in Figure 3.5.

A Task: The task, certainly, needs to be sufficiently complex to constitute a suitable benchmark for examining a given ontology. Especially if the target of the evaluation is to include non-taxonomic relations as well, it is necessary to find tasks where the performance outcome hinges substantially on the way these relations are modeled within the ontology.

One (or more) Ontologies: This almost goes without saying, at least one ontology is needed for the type of evaluation proposed herein. However, note that one is sufficient, i.e. as an ontology is evaluated in terms of its fitness for a given task, this can be done as a single ontology evaluation as well as an evaluation of how one ontology fares on the specific task as compared to another. It bases, therefore, in principle on the same paradigm as applied in the TREC, MUC or SENSEVAL evaluations.

An Application: As an application one specifies the specific algorithm that uses the ontology to perform the task at hand. To foreshadow, in part, the conclusion of this experiment, the untangling of algorithmic and ontology-related factors constitute the most difficult issue in this approach and it is vital that the algorithmic side is kept constant within an evaluation suite.

A Gold-Standard: In order to evaluate the performance of any algorithm that produces so-called *keys*, whether they be part-of-speech tags, word senses or extracted ontological relations, a given set of *answers* is needed. I have referred to this perfectly annotated solution or corpus of answers a *gold-standard*.

3.6.2 The Data: An Evaluation Suite

In the following, I describe how this evaluation framework for measuring ontological fitness in terms of the proposed error rated is instantiated in Task_F.

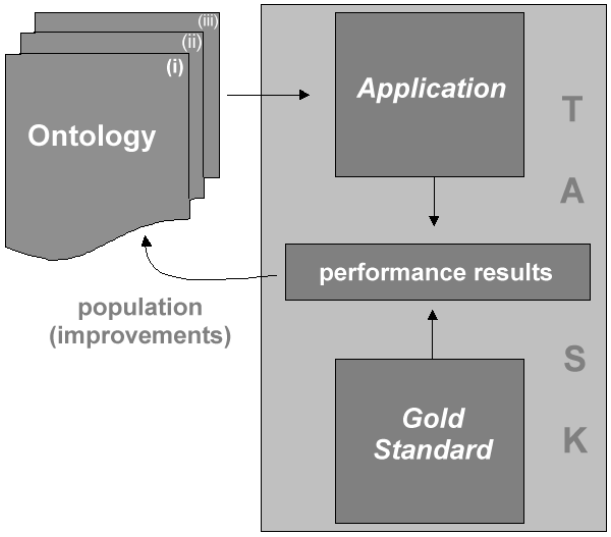

Figure 3.5: Test Suite Setup with a single task, an application, a gold-standard and one or more ontologies

A Sample Task: For this trial evaluation suite, I will refer back to the task of extracting the *semantic relations* that hold between nodes of the ontology. This mark-up can be gained from the relation extraction system described in Section 3.5.1 and constitutes a form of semantic labeling system, whereby the specific *markables* correspond to items from the ontology's inventory of semantic relations. As described above this task can be thought of as an extension of the work by Gildea and Jurafsky (2002), wherein the tagset is defined by entities corresponding to the annotated FrameNet corpus [Baker et al., 1998]. Additionally, the task discussed herein features similarities to the *scenario template task* of the Message Understanding Conferences [Marsh and Perzanowski, 1999]. In this case predefined statements are given - as shown in Example 25 in Section 3.5.1. Again, the task of concept has to be considered solved successfully, i.e. all lexical items that have ambiguous word-to-concept mappings, such as given in Example 24 in Section 3.4.3 have been disambiguated correctly. In this experimental suite I can, therefore, employ the specific Task_E of semantic relation extraction - i.e. to label all previously disambiguated and concept-tagged words with non-taxonomic relations, such as shown in Figure 3.28 - for this more general Task_F.

A Sample Ontology: The ontology used is this experiment is the one described in Section 3.1.3. Note that the hierarchy of semantic relations aligns with the frame semantic analysis used in the FrameNet project [Baker et al., 1998]. The taxonomic structure of the semantic relations itself also reflects the general

Table 3.28: Extracting ontological relations *has-channel* and *has-broadcast* for the set of concepts *Broadcast*, *Channel*, and *RecordTapeDevice*

	Concept Set	Score of Formula 3.4	Concept-Word Ratio
C_1	*Broadcast* *Channel* *RecordTapeDevice*	$S\prime(C) = 0.81$	$\frac{C}{W} = 0.5$
	Concept	Relation	Concept
	r_i: *Broadcast*	*has-channel*	*Channel*
	r_j: *Channel*	*has-broadcast*	*Broadcast*
	r_k: *RecordTapeDevice*	*has-broadcast*	*Broadcast*
	r_l: *RecordTapeDevice*	*has-broadcast*	*Broadcast*
	r_l: *Broadcast*	*has-channel*	*Channel*
	Concept Set	Score of Formula 3.4	Concept-Word Ratio
C_2	*Broadcast* *Channel* *RecordTapeDevice* *TwoPointRelation*	$S\prime(C) = 0.43$	$\frac{C}{W} = 0.67$
	...	...	...

intention to keep abstract and concrete elements apart. A set of most general properties has been defined with regard to the role an object can play in a process: *has-agent*, *has-theme*, *has-experiencer*, *has-instrument* (or *has-means*), *has-location*, *has-source*, *has-target*, *has-path*. These general roles applied to concrete processes may also have subslots: thus an agent in a process of buying as a *TransactionProcess* is a *buyer*, the one in the process of cognition is a *cognizer*. This way, slots can also build hierarchical trees. The property *has-theme* in the process of information search is a required *has-piece-of-information*, in presentation process it is a *has-presentable-object*, i.e., the item that is to be presented.

A Sample Application: The performance of relation extraction system described in Section3.5.1 depends on the given ontological model as its representation of domain context, which is employed as described in Section 3.2.3 using Formula 3.4. The input was constituted by n-best lists of speech recognition hypotheses from the SmartKom system [Wahlster et al., 2001] computed out of the ASR word graphs [Engel, 2002] as described in Section 3.5. As described beforehand, it was evaluated successfully on a number of tasks, i.e. Tasks A through D for computing a numerical ranking of alternative SRHs and thus providing an aid to the task spoken language understanding, by resolving noise and ambiguities. More precisely, the tasks have been to evaluate the best SRH suitable for further processing, as discussed in Section 3.3, or the best concept mapping, examined in Section 3.4, it in terms of its context-dependent representation within the domain and discourse model

Please also note again the distinction between the kinds of direct relations

- that can connect two nodes (concepts) in ontological models - i.e. *isa* and semantic relations. The weights of the direct relation in the underlying algorithm - described in Section 3.2.3 - are 0 for *isa* relation is set to an 1 for semantic relations. The algorithm selects from the set of all paths between two concepts the one with the smallest weight, i.e. the *cheapest*. Thereby determining a semantic relation chain between all concept pairs in a given conceptual contextual representations, excluding those that are solely connected via the *isa* hierarchy and scored with the maximal distance D_MAX.

A Sample Gold-Standard: For this I can employed the annotation-based gold standard of concept tagged data set consisting of speech recognition hypotheses that had already been identified as being the best ones. For these utterance representations the ontological relations that hold between the concepts that are part of the ontology's process hierarchy and the concepts that are part of the ontology's physical object hierarchy had to be identified.

As this is quite a difficult task and requires substantial knowledge of both the relation inventory and its semantics, as described in Section 3.5 two annotators were trained for this task to examine if their inter-annotator agreement was sufficient to conclude that this is a task that human annotators can reliably undertake. The resulting inter-annotator agreement on this task amounted to 79.54% as shown in Table 3.23. This shows that the relation tagging task is executable by humans with a satisfying degree of reliability. The corresponding gold-standard was, again, produced by means of the annotators agreeing on mutually satisfactory solutions for the cases of disagreement.

3.6.3 The Results: Ontological Fitness

For evaluating the fitness of a given domain model I proposed and described the semantic relation error types listed in Table 3.27 above. Also, I defined a correctly identified relation if the non-taxonomic relation chosen was labeled as accurate. Inaccurate ones featuring these relational errors, which are manifested either by:

- deletions, i.e. missing relations in places where - according to the annotators - a relation ought to have been identified,
- insertions, i.e. postulating any relation to hold where none ought to have been, or
- substitutions, i.e. postulating a specific relation to hold where some other ought to have been.

An example of a substitution in this task is given with the corresponding utterance in Example 26. Again, in this case the concept disambiguation was accurate, so that the two ambiguous entities, i.e. *kommen* and *Schloss*, were correctly mapped unto a *MotionDirected* process and a *Sight* object - the concept *Person* resulted from an unambiguous word to concept mapping from the form

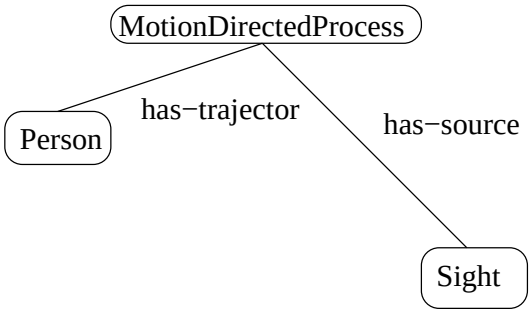

Figure 3.6: Substitution Type A: The gold-standard relation *has-target* was substituted with the relation *has-source*

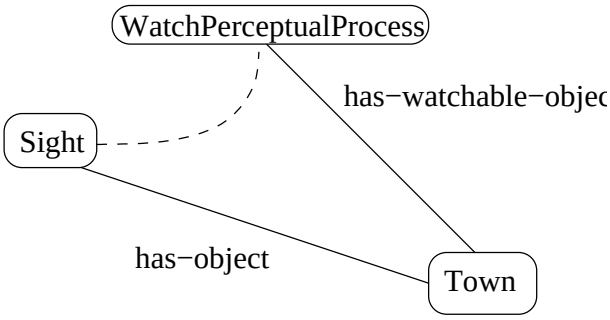

Figure 3.7: Substitution Type B: The gold-standard relation *has-watchable-object* was linked indirectly via the concept *Town* with the relation *has-object*

ich (*I*). An extracted example of relations for this case is given in Figure 3.6. In this case the relation *has-source* does not fit to the gold-standard one *has-goal*. This, of course, is due to missing syntactic and word information.

Those cases shown in Figure 3.7 (Type B) also were counted as a special case of substitution, they accounted for about 50% of all substitution errors as discussed in Section 3.5.3. The gold-standard produced by human annotation, as discussed in Section 3.5.2, therefore, contained cases as inaccurate where a *relation chain* was selected by the algorithm instead of a direct relation. These cases, such as the connection between *WatchPerceptualProcess* and *Sight* shown in Figure 3.7, were considered substitution errors, because a direct relation was indicated as a substitution in the gold-standard.

As a deletion such cases were counted in which the gold-standard containing a specific relation - such as *WatchPerceptualProcess has-watchable-object Sight* - was not tagged at all by the system, as shown in Figure 3.8. On the opposite side an insertion was counted where any relation, e.g. between *Agent* and *Sight* in Figure 3.9, was tagged by the system.

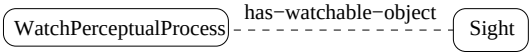

Figure 3.8: Deletion: The gold-standard relation *has-watchable-object* was not tagged by the system

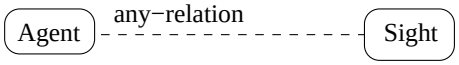

Figure 3.9: Insertion: Any relation was tagged where gold-standard

An Evaluation of the Evaluation Experiment

An overview of the percentages of substitutions, deletions and insertions of all relations extracted is shown in Table 3.24. Please note again, that the relations under consideration were extracted by the system described in Section 3.5.3, which employs these *semantic paths* for scoring the distance between two nodes given a formal and explicit domain context model. A closer examination shows that not every error can be mapped directly for improving the fitness of the ontological model at hand. For example, in cases of substitutions of Type A the appropriate conceptual instrument exists in the model. In this case, either role *has-target* and *has-source* was available and could have been chosen. The selection of the erroneous relation, therefore, was caused by the application/algorithm and not due to a shortcoming in the ontological model.

All substitutions of type B as well as deletions can, theoretically, be used for populating the ontological model with new or better instruments. The term *instruments*, as I use it here, denotes that not only concepts are/can be added but also semantic relations that are missing or modeled inefficiently. This, however, is not to say that each corresponding change automatically leads to an improvement in fitness, i.e. the specific task-performance. In my mind, a tenable expectancy for corresponding experiments is that a specific percentage of the corresponding changes in the domain model will uncover new errors, which will have to be examined - employing the paradigm proposed above - again. In theory, his can lead to an iterative quantitative evaluation approach not only for estimating the fitness of domain models employed as contextual representations, but for other tasks as well. Alternative applications for such a task-based evaluation are constituted by sense tagging and discovering set-ups or learning experiments [Pantel and Lin, 2003]. While these can lead to concept population or concept generalizations [Widdows, 2003b], the task of evaluating and improving an ontology's non-taxonomic relations has been out of their scope.

The population fall-out of our framework, in this case, derives its content from the gold-standard that was merged from the doubly annotated data. This manual input is still required and without it the task of improving the *se-*

Table 3.29: Results Ontological Fitness Experiment

fitness task on corpus $REL_{1'}$	substitution errors	deletion errors	insertion errors
Fitness	15.32%	7.11%	1.26%

mantic relations automatically is still an open challenge for ontology learning approaches. The result of this evaluation makes it more feasible to measure our progress along the path to better performances and better ontological models as they clearly indicate several shortcomings in the ontology used:

- The 7.11% deletions indicate clear cases where a pertinent (at least for this task) relation was not modeled in the ontology,
- about 50% of the substitution errors showed inefficiencies in the model (the rest were a result of the algorithm's shortcomings), and
- the - however - small percentage of insertions can be regarded as superfluously modeled relations.

It would now be possible to go back to the model and undertake the corresponding changes and run the evaluation again iterating this process until the accuracy approaches a desired value. While this *optimization* of an ontology for a given task is not within the scope of this work, it is nonetheless important obtain validated explicit context models that are fit for the task. Further challenges and research questions that arise with this a task-based approach to ontology evaluation concern ways to make the proposed framework more general and scalable [Porzel and Malaka, 2005, Gangemi et al., 2005]. Nonetheless, this examination of the fitness of this specific contextual representation, will be considered as concluded with the results obtained and discussed above. Following a final look at the contributions achieved *vis à vis* the work presented in this chapter, I will turn to face the remaining two context types and the final task type from the list given in Table 3.1 at the end of Section 3.1.5.

3.7 Summing-up

In the examinations I have presented in this chapter formal and explicit representations of domain- and discourse-context were employed on a number of tasks, specifically those listed below:

Task Accurate - classification of speech recognitions hypotheses in terms of their semantic accuracy;

Task BestOf - classification of the best speech recognitions hypothesis from a set;

Table 3.30: Overview of Domain and Discourse Context Results

Task	majority class baseline precision	contextual computing precision	gain over baseline	relative human precision
Accurate	≈ .52	≈ .65	.13	≈ .80
cum discourse	≈ .52	≈ .66	.14	≈ .80
BestOf	≈ .64	≈ .84	.20	≈ .95
cum discourse	≈ .64	≈ .88	.24	≈ .95
Coherent	≈ .63	≈ .70	.07	≈ .80
cum discourse	≈ .63	≈ .71	.08	≈ .80
Disambiguation	≈ .52	≈ .64	.12	≈ .79
cum weights	≈ .52	≈ .65	.13	≈ .79
Extraction	n.a.	≈ .76	n.a.	≈ .80
Fitness	≈ .15	≈ .07	≈ .01	
(error type)	(substitutions)	(deletions)	(insertions)	

Task Coherence - classification of speech recognitions hypotheses in terms of their internal coherence;

Task Disambiguation - classification of word sense by resolving multiple word to concept mappings;

Task Extraction - extracting contextually appropriate semantic relations for individual concept pairs;

Task Fitness - evaluating the context representation itself.

The underlying contextual computing approach has been based on ontology-based representations of domain and discourse context that are evoked by given processed spoken input. In the task-specific examinations of the contribution of adding such explicit context models to help understanding what was said and meant significant gains where achieved over the task-specific majority class baselines in Tasks A through D. An aggregate overview of all results reported in this chapter is provided in Table 3.30. Where applicable, I also presented a calculation of the probability of these gains having arisen by chance, i.e. their statistical significance - calculated by means of a corresponding unpaired t-test results in $p = 0.011$ for Tasks A through C and $p = 0.003$ for Task A through D.

The results of this contextual computing approach thereby confirm the value of identifying the semantic relations that hold between the entities evoked by the linguistic forms for understanding their meaning. Consequently, an understanding of the roles which the entities play in a given context goes a long way towards determining corresponding semantic specifications, which, therefore,

constitutes the core task in analyzing the meaning of a natural language utterance [Chang et al., 2002, Bryant, 2003, Feldman, 2006]. The results gained also show - more specifically - that this knowledge can also assist in increasing robustness against noise and in cases of semantic ambiguities.

However, although it can be seen as quite an achievement to approximate the output of a *semantic parser* to such an extend, obvious shortcomings are caused by missing linguistic information. Including this information for finding the best-fitting semantic specification has also been successfully performed on a corpus of spoken child-directed utterances in Mandarin Chinese [Bryant, 2008]. Therein, language-specific constructional forms are analyzed to specify the roles of explicit and implicit entities against the backdrop of explicitly modeled conceptual schema as discussed in Section 2.7.2. Given recent modeling instruments, that will be discussed in the next chapter, it is also feasible to incorporate both linguistic information and the ensuing conceptualizations in ontological models in both more traditional morpho-syntactic approaches or constructional ones [Buitelaar et al., 2006, Porzel et al., 2006b].

Before concluding this chapter, I want to point out that the system described above in Section 3.2.3 has been implemented also in a context modeling component which is employed by the SmartKom multi-domain spoken dialogue prototype. There it is applied for the task of scoring n-best lists of alternative noisy and ambiguous utterance representations of spoken utterances, thus producing a score expressing how well the evoked conceptual sub-graph fits with respect to the given domain and discourse context. Furthermore, such representations - based on formal ontologies - will be reintroduced in Chapter 4.2 in light of the completely ontology driven SmartWeb system [Reithinger et al., 2005, Cimiano et al., 2004].

As noted in Section 1 an additional challenge for natural language processing - in addition to noise and ambiguities - concerns underspecification. This problem of underspecification goes beyond the field of traditional semantics into the domain of pragmatic interpretation as discussed in Chapter 2. A prototype example being that of so-called *conversational implicatures*, where the *function* of the utterance is more implied in that explicated in the utterance. As will be discussed in greater detail below, this problem domain features some notable differences. One methodological difference, for example, lies in available annotated data and ensuing baseline computations. I will return to the question of *how to annotate what is not there* in Section 4.2. However, underspecification, which occurs frequently in unrestricted dialogues, is at hand whenever some implicit information needs to be explicated in order to draw the necessary inferences. The information left implicit can, however, and must be recoverable by recourse to context. That is not to say that contextual observation alone suffices, but that they are certainly needed to run corresponding stochastic simulations or set the values in observation nodes of graphical models or as instances of the corresponding pragmatic ontological models as will be discussed below in Section 4.3.

It is important to note at this point, that this implicit information was provided the by the specific speaker in a particular situation. Via these context

types, i.e. the interlocutor and the situation as the remainder of the *modality of context* - information can be provided, for example, about the actual geographic position of the speaker, i.e. being at a given place at a certain time. While changing the focus from domain and discourse knowledge to that concerning the interlocutor and the situation, the question and research focus remain on assessing the benefits of taking context into account. I will assume a perspective, encompassing the whole modality of context, in the final discussion of the contextual computing approach presented herein in Chapter 5, but for now invite the reader to switch their focus on more pragmatic matters.

3.7.1 Roadmap

The classification, extraction and evaluation experiments presented in this chapter showed how contextual computing can be performed by recourse to ontological representations of domain & discourse knowledge and evaluated the ensuing performance several tasks followed by an evaluation of fitness of the central semantic resource employed in these examinations. Now, as the road approaches the interlocutor and situation at hand the scope also broadens as it inclines towards the field of pragmatics. There, I will take a closer look at the specific types of knowledge concerning the speaker and the situation that have been examined in the past in Section 4.1. I will then examine the case of resolving pragmatic ambiguities in natural language understanding in Section 4.2, before presenting the corresponding formalization thereof in Section 4.3.

Chapter 4

User and Situation

Studies on context in human language were not predominant in the influential research direction that bases on an assumed *autonomy of syntax* [Chomsky, 1965] and that explicitly excluded *performance* related issues as out of scope of their analysis [Chomsky, 1981]. These analyses focused on linguistic competence as well as innate structures and -mechanisms for the acquisition of such a (universal or core) grammar [Chomsky, 1995]. Noting the omission of semantic - let alone pragmatic or contextual - considerations by the so-called *East-Coast* linguistic school, the so-called *West-Coast* school proposed an alternative point of view [Langacker, 1987] that flourished under several headings, e.g. *cognitive grammar* or *functional grammar* [Givón, 1995], but is essentially usage-based. Here linguists started to take into account that actual utterances are addressed at someone - the interlocutor(s) and that they actually happen in real situations.[1]

The following sections I will, therefore, discuss how contextual factors that concern properties of the interlocutors and the situation critically influence spoken language understanding and production. As before, I will present and employ empirical data and methods to examine their influences on the processing of natural language utterances embedded in this approach to contextual computing and its corresponding methodological framework. As stated in Section 1.1 speakers may not always be aware of the potential ambiguities and underspecifications inherent in their utterances. They leave it to the context to disambiguate and specify the message, i.e. to decontextualize in the sense of McCarthy or, to put it linguistically, to resolve ambiguities and to specify elided information. Speakers must, therefore, trust to some degree in the addressee's ability to perform such context-specific *leap* from the utterance to arrive at the illocutionary function that they wanted to elicit [Katz, 1980]. In order to interpret context-dependent utterances correctly the interlocutionary partners must also share - or at least have access to - the same interlocutionary and situational context as well as to the domain and discourse context discussed in Sections 3.1

[1] A little of the surprise diminishes when considering that decades of so-called *armchair-linguistics* have dealt with written sentences self-assembled by the respective linguist in his or her armchair and that real situations and dialogs are difficult to create in laboratory conditions.

through 3.6.

4.1 Modeling User and Situation

For explicit contextual computing approaches this *sharing* of context entails access to several knowledge models as well as corresponding information regarding the actual instances thereof as described in Table 2.7. Moreover, the need to include this set of contextual knowledge stores for natural language processing increases as speakers anticipate the employment of these interpretative resources[Branigan and Pearson, 2006]. It is, furthermore, economic and increases dialogical efficiency when speakers as well as systems - anticipate the employment of these resources and construct the utterance knowing that certain underspecifications as well as other forms of alignment are possible, since the hearer can infer the omitted information or can resolve referents despite ambiguities faster then it takes to explicate them.

Generally speaking, this critical anticipation of the interpretative resources of the dialog partner - whether it be a human or an artificial interlocutor - is based on the speaker's mental model of the dialog partner. I employ the term *mental model* here in the more general `human-computer interaction`-sense of Norman (1988) [Norman, 1988]. The process of tailoring any form of linguistic behavior or output towards the recipient of that output has been labeled variously as *listener-*, *user-* or *partner-modeling* [Levelt, 1989, Paris, 1993, Glatz et al., 1995].[2]. I will, therefore provide an corresponding overview of pertinent and representative findings and approaches first for modeling the interlocutor in Section 4.1.1 and then for modeling the situation in Section 4.1.2.

4.1.1 Modeling the User

The modeling of interlocutionary context has been the subject in various areas of scientific scrutiny, for example in textual semiotics this context type has been labeled as *the role of the reader* [Eco, 1984] and in socio-linguistics as *the role of the listener* [Krauss, 1987]. In artificial intelligence research, e.g. in the area of *intelligent user interfaces* [Maybury and Wahlster, 1997], *user modeling* has been the predominant heading under which research has been performed in order to pave the way towards more user-adaptive interfaces. Therein, adaptive models of the user have been examined and employed for various purposes. In work on multimodal systems user and situation models have been employed for modality fission and presentation management [André, 1999, Elting, 2002, Reithinger et al., 2003]. Note that for multimodal systems - as discussed in Section 2.4.2 - modality fission constitutes the *output pipeline* that determines the modality-specific forms to be employed in conversing with the user.

[2]Computationally, the term *User Modeling* has traditionally been employed, e.g. in the special issue on User Modeling published by the Journal of *Computational Linguistics* (Volume *14*, 3) already in 1988.

Congruently, in the area of natural language processing interlocutionary context - as user modeling - has been implemented in natural language generation systems [Jameson and Wahlster, 1982, Paris, 1993, Bateman and Zock, 2003]. These systems include user-models as well as dialog-specific discourse models and representations of the domain knowledge as discussed in Section 3.1. On the other side of the coin specific design choices can be used to evoke specific linguistic behavior by the user [Branigan and Pearson, 2006]. Their common goal is to enhance the context-adaptive capabilities of these systems, e.g. to generate or elicit appropriate linguistic expressions for or from different users by employing interlocutionary models.[3]

Regardless of the terminology employed the proposed interlocutionary models all assume domain-specific knowledge, which I have discussed in Chapter 3, empirical findings both suggest modality-specific preferences [Elting et al., 2002] and register-specific preferences[Fischer, 2006] specific to the interaction with artificial multimodal systems. Such situation-specific and interlocutor-specific preferences, therefore, constitute additional contextual factors to be considered. Such empirically derived data - in a sense - spells out the content of the often vague and arbitrary employed category of *user preferences* that populates most models referenced above. Most models, e.g., the one employed by Paris (1993), assume knowledge of the user's goals and plans as well other contexts such as beliefs, interests and numerous physical attributes of the user (Ibid:17f.). Nevertheless, adaptation to the user has been implemented and formalized for modeling goals and plans [Anderson et al., 1995], epistemic factors, such as prior knowledge [Paris, 1993], and for situational factors for generation of multimodal output based on models of user-dependent multimodal preferences [Elting, 2002].

In general, context-dependent selection, composition and construction of information has been implemented where the adaptation and alignment hinges on knowledge about the interlocutor. The corresponding (interlocutionary) context models have been discussed as *listener-*, *user-* or *partner models* and are of central importance for intelligent multimodal interaction. Moreover, empirical studies, as, for example, conducted by linguistic research [Fischer, 2006, Branigan and Pearson, 2006], can shed some light on how these mental models of our interlocutors are constructed in context-specific conversations.

A substantial amount of empirical work exist for the specific context of spatial language, employing various types of utterances within the domain of space, such as instructions - e.g., in the form of spatial directions - and localizations - e.g., in the form of descriptions - which I will also discuss for the experiments presented in the following Section 4.2. Beforehand, I will present the prior work on models of interlocutionary and situational context in the domain of spatial language as specific forms of user-specific alignments, e.g. the empirical studies performed on spatial perspective taking in conversations [Schober, 1993, Herrmann and Grabowski, 1994]. The general necessity of the

[3]The ensuing generation problems includes all the referential problems inversely, e.g. when to employ an anaphora or produce an ellipses [Jameson and Wahlster, 1982, Strube and Wolters, 2000].

inclusion of interlocutionary context has been discussed before and seems undisputed at the moment, since without recourse to the contextual knowledge store of a partner model several well known empirically observable phenomena cannot be explained.

Already in 1987 socio-linguists, such as Robert Krauss, broke with traditional 'context-free' sender-receiver models by stating that:

> the traditional separation of the roles of participants in verbal communication into sender and receiver, speaker and addressee, is based on an illusion — namely that the message somehow *belongs to* the speaker, that he or she is exclusively responsible for having generated it, and that the addressee is more-or-less a passive spectator to the event. I am not denying that the speaker is responsible for the physical act (...). But (...) the addressee is a full participant in the formulation of the message — that is the vehicle by which the message is conveyed — and, indeed, may be regarded in a very real sense as a cause of the message [Krauss, 1987]:96

The interlocutor subsequently has been regarded in usage-based empirical approaches to play an essential part in the causation of speech production in a dialogical setting. Moreover, to ignore the interlocutionary context during the execution of a dialog can be costly in terms of dialogical efficiency and task completion. Studies on the dynamics of alignment in dialog started with work on *back-channeling*[4], which labels a linguistic side of the multimodal alignment phenomenon much as the work on *entrainment* [Garrod and Anderson, 1987, Brennan, 1996, Brennan, 2000] constitutes another side of linguistic alignment. In a multimodal light, gestural and other non-linguistic forms of back-channeling a form of alignment that has been examined under the heading of *behavioral mimicry* including mimics and body movements have been examined in human-human settings [Sweetser, 2003, Sebanz et al., 2006] as well as in interactions with artificial agents [Kopp et al., 2004, Kraemer et al., 2007].

The effects of alignment and non-alignment to the interlocutor have been examined in the field of human-human conversation showing that dialogical efficiency can be influenced by back-channeling which either enhanced efficiency by reducing redundancies of words and phrases or decreased efficiency by causing lexical and phrasal repetitions [Krauss and Weinheimer, 1964]. In the event of back-channeling also more economical shorthands, e.g. abbreviations and phrase-reductions, become employed alongside other forms of entrainment. Visual back-channeling also increases the efficiency of the discourse [Krauss et al., 1977]. The efficiency-effects of dialog-structuring particles on turn-taking strategies in human-human interaction have also been examined thoroughly [Duncan, 1974, Sack et al., 1974, Weinhammer and Rabold, 2003]. More specific findings for (non-)alignment comes from psycholinguistic research

[4]This term denotes verbal and para-verbal responses of the listener [Yngve, 1970], which occur during the dialog manifested by specific linguistic forms such as *yes*, *hmmm*, *I see*, *uh-huh*.

on perspective-taking in the spatial domain, e.g. work on the production of spatial descriptions [Schober, 1993, Herrmann and Grabowski, 1994]. In the latter experiments the principle methodology employed was to create situated interaction keeping the domain context - a spatial state of affairs - and discourse context - a given communicative task - constant while changing interlocutory and situational context; where they varied the interlocutors position, social status, assumed cognitive competence. For the most frequent case, i.e. perspectival non-alignment, additional research on perspective-taking in the domain of spatial descriptions has demonstrated that localizations aligned to the interlocutors position in space demand more cognitive resources than non-aligned (egocentric) localizations [Bürkle, 1986], which is congruent to empirical findings concerning cognitive efforts in mental rotation tasks [Shepard and Metzler, 1971, Shepard, 1975]. Please note, that in another context, i.e., the one created by Branigan and Pearson (2006), lexical alignment constituted the default and lower levels of entrainment were induced by changing interlocutionary context in the interaction with an artificial dialog system.

Looking at the computational side only some prior work exists concerning the turn-taking strategies of dialogue systems in human-computer interaction, e.g., for the case of conversational computer-mediated communication aids for the speech and hearing impaired [Woodburn et al., 1991] or for turn negotiation in text-based dialogue systems [Shankar et al., 2000]. It has been noted before, that that problems, such as turn-overtaking, -handling and -repairs, have not been addressed by the research community [Wooffitt et al., 1997]. Also in the context of the research performed in the SmartKom context, studies show both the drastic effects of ignoring the interlocutors turn-taking signals [Beringer, 2003] as well as specific effects of these interlocutionary signals on dialogical efficiency [Porzel and Baudis, 2004].

There is also a general discussion about the status of the discourse context in regards to the interlocutionary context. As pointed out in Section 3.3 the discourse model contains the discourse-protocol or -history, i.e., at least a representation of the referents already introduced into the discourse and the statements made about them. The ensuing discussion deals with the problem that, when a statement about something has been made, say proposition p about the domain d has been uttered, then p becomes part of the user's domain knowledge about d and needs to be included in the information contained within the user model as *common ground*. The influence of common ground, i.e., the shared knowledge, shared associations, shared sentiments, and shared defaults, between speaker and listener has been identified before [Kingsbury, 1968, Krauss et al., 1977, Clark and Marshall, 1981] The amount of common ground influences the lexicalizations preferred by the speaker, for example what kind of words to use, whether to describe objects more figuratively or literally. Furthermore, it influences the type versus token ratio in the speakers' discourse.[5]

[5]In experiments tailored towards the identification of contextual-dependencies also negative findings are equally important for finding, for example, that type-token ratio is not influenced by interlocutionary information supplied via the aforementioned feedback channels [Porzel and Baudis, 2004].

Clark and Marshall found also that length and specificity of descriptions are demonstrably influenced by common ground between the interlocutors, which raises further research questions for conversational interfaces to artificial systems [Wilson, 1997, Cassell, 2001, Shneiderman and Plaisant, 2004, Dourish, 2007].

Generally, effects of the user's estimations concerning properties of the system influence formalization greatly, which is known from examining the characteristics of computer-directed language [Zoeppritz, 1985, Wooffitt et al., 1997, Darves and Oviatt, 2002]. More recently, dedicated empirical studies on alignment and non-alignment in language and other modalities have been performed [Garrod and Anderson, 1987, Brennan, 1996, Brennan, 2000], Additionally, it has been shown that context-dependent constraints can be imposed by the situation, e.g. lawyers choosing not to entrain on critical terms [Brennan, 1998], or that subjects can be induced to adapt their amount of entrainment depending on the assumed features of the computer system [Branigan and Pearson, 2006]. Again, this exemplifies how interlocutory and situational context affect alignment depending on interlocutionary perspectives [Fischer, 2006][6] as well as on the constructed *mental models* [Norman, 1988].

Research teams have increasingly mounted evidences that lexical and phrasal, i.e. constructional, choice within a dialogue is dependent also on the epistemic stance adopted, which is seen as interlocutor-specific and, therefore, context-dependent. To adopt the interlocutors perspective - e.g. through linguistic hedging - two interlocutors adopt each other's terms. The variability - counterparting the ambiguity discussed in Section 3.4- in constructional choice is huge in any domain. On the lexical level, it has been called the *vocabulary problem* [Furnas et al., 1987]. Although usage-based corpora clearly show there are no real synonyms, i.e. two words that in all contexts could be used interchangeably, speakers still have numerous context-dependent options when referring to an object, i.e. to find a form for a given meaning. For instance, in the user study conducted by Furnas *et al.* different subjects used *delete*, *change*, *remove*, *spell* or *make into* to denote the same, situationally given, event.

Another acquisition bottleneck exists for systems seeking to adapt their *behavior* to their users. They must be provided with the means for acquiring corresponding user models. A description of the problem is provided by Chin (1993), who also provides an overview of approaches to circumvent the *invasiveness problem* [Chin, 1993]. Generally, dialog systems score higher in user satisfaction measures, which are able to involve user-specific adaptation by means of direct questioning or observation throughout a session [Walker et al., 2000].

Before moving to models of the situation, I want to note that interlocutionary context encompasses both interlocutors, e.g. a human user and an artificial agent. The model of the artificial system itself, i.e. the system's topical self model, is important as well. The given representation of the system's state can influence, for example, via dedicated profiles, the output generated by the system can depend on a given device configuration, network bandwidth or compu-

[6] This can be seen to constitute a linguistic analogue of the classic framing problem [McCarthy and Hayes, 1969].

tational load, as in *quality of service* approaches [Mammeri, 2004]. For the case of multimodal systems, system models have been employed for device-dependent modality fission and presentation management [Malaka et al., 2006].

4.1.2 Modeling the Situation

On the one hand - looking back at the models of domain-, discourse- as well as the empirical findings regarding the interlocutionary context, that I presented above - it is notable that for natural language processing systems, e.g. those presented in Section 2.4.1, prior research on contextual computing more or less excluded situational context. In earlier dialog systems, such as TRAINS or TRIPS, the *situation* was given as a hypothetical usage scenario and, therefore, in a sense *hard-coded* implicitly into the system, consequently, the system's behavior was not effected by the actual location where it was used or demonstrated or how the real weather was at the time. Certainly, hardware issues, such as desktop size, processing resources and power supply restricted earlier dialog systems to stationary applications that were exhibited in enclosed laboratory, conference or trade-fair settings.

As I have noted in my introductionary remarks, the advent of mobile computing brought forth first mobile prototypes in domains such as tourism, geographic information systems and other more-specific location-based services, e.g. involving hotel, restaurant or cinema reservation systems [Johnson, 1998, Johnston et al., 2002, Malaka and Porzel, 2000, Wahlster et al., 2001]. Since then, mobile *incarnations* of dialog systems have added numerous domains which was facilitated on the one hand by coupling ontological representations, such as the ones described herein, to the morpho-syntax of web service descriptions [Oberle et al., 2005]. This enables mobile multimodal systems, such as the SmartWeb system to access information regarding the topical situation at hand in their own ontological vocabulary [Wahlster, 2004], as I will discuss below in Section 4.2.

Research on multimodal systems consequently included mobile scenarios involving the spatial domain as a suitably complex challenge for an intuitive conversational natural language processing system, as described in my discussion of the state of the art in Section 2.3. Employing these situation-aware systems, mobile users can ask for directions or localizations of places or sights [Johnston et al., 2002, Malaka and Porzel, 2000, Wahlster et al., 2001] using deictic expressions such as *from here*, which will be resolved and grounded by the location-aware natural language processing system. The resulting prototypes - i.e. mobile tourist information systems that guide users through cities, can provide detailed spatial, architectural and historical information as well as topical information from hotel, entertainment and weather services [Coors et al., 2000, Oberle et al., 2005].

In the field of artificial intelligence several approaches have sought to model the semantics of situations explicitly ranging from model-theoretic approaches via ones employing modal logic [Barwise and Perry, 1983] to more recent ones, which were motivated also by ontology engineering perspectives within the se-

mantic web framework [Gangemi and Mika, 2003]. As the focus of this work will return to the formal approaches pertinent to describing contextually construed situations in Section 4.3.1 and - as their formal properties and shortcomings have been introduced in Section 2.2 - I will conclude this overview by pointing at the important work on situation models in cognitive and linguistic approaches. This work ranges from finding prototype situation models [Zwaan and Radvansky, 1998] and categorizations [Rosch, 1983] to modeling the resulting cognitive prototypes [Lakoff, 1987].

4.1.3 Pragmatics in SmartKom

I have discussed in Section 2.4.1 how contextual considerations in early dialog systems were restricted to low-level processing [Bunt, 2000] and by and large excluded pragmatic analyses. Additionally, they encompassed small domains and featured a pre-defined modular processing of the spoken input that produces a formal representation of the given input as described in Section 2.3. This representation, then, becomes equated with the user's intention. Starting with the VerbMobil project the dialogical situations grew more challenging, which lead to a discourse-sensitive mapping of the parsed speech input to specific utterance types, e.g. *confirmation* or *question* [Reithinger and Maier, 1995, Alexandersson et al., 1995], that represent the set of *speech-acts* [Austin, 1962, Searle, 1975] encountered in the VerbMobil domain. Still, the VerbMobil speech-to-speech translation system featured a single domain and desktop scenario, moreover, the resolution of deictic, anaphoric or underspecified expressions was ultimately left to the other human interlocutor and the system itself did not act as a dialog partner.

Nevertheless, continuing research on dialog systems brought forth several multi-domain prototype systems that are set in a mobile computing context [Malaka and Porzel, 2000, Malaka et al., 2006, Ankolekar et al., 2006]. Consequently, first corpora of more conversational computer-directed speech in outdoor situations - ranging from pedestrian to car and even motorcycle drivers [Schiel et al., 2002, Kaiser et al., 2006, Mögele et al., 2006] - have become available. Before addressing the additional challenges encountered in going from a multi-domain system to an open domain one in Section 4.3, I will start this examination with the contextual computing tasks afforded by multi-domain systems, such as DeepMap or SmartKom, where linguistic expression are found in the data that are underspecified and ones that can be understood in multiple ways.

As noted before, the human enterprise of answering or responding to conversational speech input in a suitable and felicitous manner, it is not solely based on the ability to recognize what was said by the questioner, but also requires the ability to infer information that is left implicit by the questioner and to estimate what constitutes a useful and felicitous answer. The realization of such abilities poses a formidable challenge in the development of conversational and intuitive dialogue systems with more than one domain, modality, or situational context. The SmartKom system, for example, has to deal with contextual dependencies

as well as cross-modal references based on the system's symmetric multimodality [Wahlster, 2003]. Moreover, it has been designed to handle multiple requests in different domain contexts and features three overall scenario-specific situational contexts.[7]

Faced with underspecified utterances that leave some information implicit, some form of *decontextualization* is needed to resolve the arising contextual ambiguities [McCarthy, 1986, McCarthy, 1990]. In the case of restricted and controlled single domain systems, the problem of contextually implicit information can be solved by generating full paraphrases out of the underspecified user utterances [Ebert et al., 2001]. In systems with - in that sense - multiple contexts, such as DeepMap or SmartKom, additional knowledge sources, algorithms and dynamic contextual observations are needed as I will describe below as well as in Sections 4.2 and 4.3.

In many cases the task of resolving underspecifications and indexical expressions, e.g. spatial or temporal deixis, requires frequent recourse to both discourse and situational context. While situational observations and context are bound to one actual specific situation, which can ground deictic expressions such as *there* or *noon*, discourse context can *override* the default groundings, by establishing a prior context in which the meaning or some particular mapping is specified otherwise, e.g., as in the poetic expression *midnight's noon*, where some part of the default meaning - middle of the day - is in a sense *inhibited*, as it is overlaid with a different meaning evoked by the prior discourse context. In the SmartKom system, discourse contextual influences are handled by default unification, implemented as an *overlay* operation [Alexandersson and Becker, 2001, Löckelt et al., 2002], which employs the hierarchical schemas which were automatically created from a reductionistic ontology and, therefore, reflect the hierarchical make-up and non-taxonomic structures of the input ontology [Gurevych et al., 2003a].[8]

As I will show in the examples given below, one can find situations where multiple types of context sources, given the categorization proposed in Table 2.7.1, contribute pertinent information. Where, so to speak, the individual *contexts* as input modalities in their own regard are interwoven with each other in much the same way as the multi-modal system-directed input streams from the user, e.g. speech, gaze and gesture, can be. Such an integration of domain and discourse models together with their respective interlocutionary and situationally given contexts constitutes a formidable hurdle to be crossed in order to achieve adaptable and scalable natural language understanding systems that facilitate felicitous cooperation and intuitive conversational interaction.

Let me note once more, that in contextually restricted dialog systems, such

[7] These scenarios encompass indoor employment with a personal device, e.g. at home or in an office, employment as a public device, e.g. in a kiosk or communication both, as well as mobile usage with private handheld- and on-board car devices.

[8] The SmartKom domain model has been described in Section 3.1.3, it was additionally employed to create the interface specifications of the system's components, thereby enabling some of them to perform their operations directly on the resulting XML schema [Gurevych et al., 2003b, Porzel et al., 2003b].

as train schedule or help desk systems [Aust et al., 1995, Gorin et al., 1997], this does not constitute a problem, since conversational phenomena can be avoided by means of the appropriate dialog design [Dix et al., 2004]. In a multi-domain system that faces diverse usage contexts, e.g. at home or in a mobile context, conversational phenomena such as underspecifications and pragmatic ambiguities add to the challenge of understanding the user's often multi-modal input. This input can refer to various types of discourse objects that play different roles that depend not only on the course of events, but also on situational parameters, such as time or place, as well as domain- and discourse-related parameters, such as what is (and has been) talked about. A comprehensive understanding of naturally occurring discourse and of the often implicit questions embedded therein still has many unsolved issues in pragmatics and computational linguistics alike.[9]

In the following, I will provide some initial examples of the kind of underspecifications and implicit information encountered in the domains at hand, followed by a respective empirical and computational examination thereof in the light of the contextual computing approach introduced herein. The primary focus of this work thereby remains an explication of the contribution of including contextual observations and explicit knowledge for enabling conversational dialog systems that face multiple contexts to realize the required understanding capabilities. Considering a question, such as given in Example 28, that can be encountered in a pedestrian setting, when a user is asking for directions.

(28) How do I get to the powder tower

Looking at corresponding human-human interactions, e.g. as described in field experiments described in Section 4.2, passerby's responses to such questions are hardly ever followed by questions where and when the spatial instructions should start. More likely, immediate directions will be - and were - given if the desired object is known to the interlocutor. But, as the collected field data, presented in Section 4.2.2 shows, the felicity of spatial instructions is also dependent on contextual factors such as distance, mobility of the questioner or weather. Information concerning time or place, for example, is rarely explicated when given *default* settings, based on *common ground* [Krauss, 1987] hold. If not, however, such information is very likely to be expressed explicitly. In some cases, which are commonly labeled as *indirect speech acts* or *pragmatic ambiguities*, however, we are not only faced with implicit information, but also with implicit intentions.

In responding to such a seemingly clear instructional request, the DeepMap system, for example, upon recognizing the goal object directed the interlocutor to that point on the modeled road network, which was closest to the center of the goal object. This, however, lead to a dead-end alley, where one could approach the particular tower of the castle of Heidelberg, but neither see nor enter it. Moreover, in this case there are three different routes depending on

[9] As made evident by misunderstandings, this comprehensive understanding can be a challenge for human dialog partners as well.

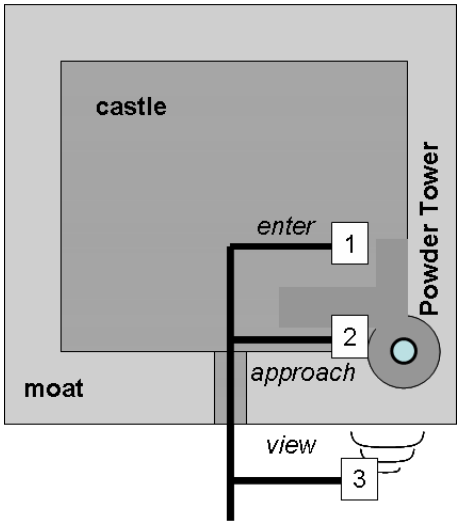

Figure 4.1: A visualization of the routes to enter (1), approach (2) or view (3) the castle tower

whether one wanted to enter, view or approach the goal object as shown in Figure 4.1. The question whether the user's goal is to enter, view or approach the mentioned goal object could hinge on a multitude of contextual factors, such as object type, accessibility or it being currently open or closed, that constitute part of the situational context. Generally speaking, due to the common ground established, a situated conversational dialogue occurring in a shared context is consequently composed of utterances based upon the pertinent knowledge of that context.

In order to find what factors are pertinent out of the diverse kinds of contextual information imaginable studies and experiments, for example of the type to be described in Section 4.2, need to be conducted to determine the individual factors and to capture their situation-specific influences formally. The task of finding out *what matters* can, in my mind, only be undertaken in the light of the specific question at hand. Looking at the domain of spatial information alone one can also imagine a multitude of additional questions that need to be posed in order to enable a dialogue system to produce felicitous responses. Next to the question, discussed above, whether the user wants to enter, view or just approach the goal object, one can ask if the user wants to take the shortest, fastest or nicest path there. Naturally, the questions regarding the mode of transportation, e.g. walking, driving, cycling or by means of public transportation, are extremely relevant to answering instructional requests felicitously.

In many cases, like the ones noted above, *solving* underspecifications corresponds to automatic context-dependent generation of more specified paraphrases. That is, to explicate the information that was left linguistically im-

plicit, e.g. to expand an utterance such as *How do I get to the castle* depending on the situational and interlocutionary context into *How do I get to the castle by bicycle on a not so steep route*. As stated before, what constitutes a felicitous answer can hinge on a number of contextual features, including also ontological factors, e.g. object type and role, situational factors, e.g. weather and environment, discourse factors, e.g. referential status, as well as interlocutionary factors, e.g. tourists or business travelers as questioners and their time constraints, may constitute significant factors.

4.1.4 Modeling Implicit Information

An attempt to model which information matters pragmatically in a given situation should firstly accommodate the empirical data, e.g. collected in situated field experiments as described in the following section. As I will show hereafter, based on such data the pertinence of the recorded factors can be assessed. In many cases, it will be necessary to combine observations from multiple heterogeneous information sources, which are not always perfectly accurate or failsafe.[10]

Based on empirical studies and implemented in a robust manner one can differentiate the following contextual computing tasks:

- to classify the situationally observed information with respect to the pragmatically relevant knowledge modeled therein, which I will discuss further in Section 4.3;
- to evaluate competing semantic specifications - as potential intentions or *intention-hypotheses* - using appropriate models of the pertinent pragmatic knowledge, which will be described in Section 4.2;
- to augment individual semantic specifications with hitherto implicit information, i.e., to spell out the underlying context-dependent underspecifications, which I will discuss subsequently.

Summarizing, an implemented context model can be employed in the classification, evaluation and explication of situationally implicit information. In order to align these contextual computing tasks with the ones discussed in the previous chapter, I will sketch out the basic correspondences in the following:

- In much the same vain as the domain model discussed in Section 3.1 provides the vocabulary of terms to classify the modality-specific input, given as instances of the linguistic- or gestural forms at hand, a pragmatic model can provide vocabulary for classifying the user- and situation-specific *input* given as instances of the situation and interlocutor.[11] This classification

[10] Robustness against missing and uncertain information constitute additional constraints that need to be heeded, as some contextual features may not always be observable, e.g. in case specific information streams of the system such as positioning sensors or external data servers are currently offline.

[11] Please note, that throughout this work, I propose to view context-specific information sources to constitute a modalities in their own right, requiring both a recognition of the context-specific input as well as an understanding thereof.

task, therefore, comes with the analogous engineering challenges and acquisition bottlenecks as one finds in the previous approach that requires the construction of a shared conceptualization of the given domain.

- In many respects the task of evaluating different intention hypothesis corresponds to that of ranking speech recognitions hypothesis, as described in Section 3.2. In both cases external models of contextually pertinent knowledge are employed to select the contextually best-fitting semantic- or *pragmatic* specifications. While the formerly presented experiments sought to elicit the contribution of the pertinent domain- and discourse knowledge, the ones to be presented shortly, do so for pertinent pragmatic- and social knowledge.

- The remaining task of augmenting a given semantic representation - in the approach proposed herein - corresponds in several respects to the relation extraction task described in Section 3.5. As described therein, frame semantic relations hold between frame elements specific to the given course of events. A correct identification of the semantic frame can, in some cases, contain frame elements that have hitherto been left implicit, but have been associated to it and the corresponding domain model.

The latter explication task concerns the provision of knowledge specifying what *belongs* to a given situation, but was omitted as a result of a shared context. Let me emphasize, once more, that - the omission notwithstanding - the associated knowledge must matter for the task at hand. It is somewhat obvious, that one constantly omits an infinite number of things with every statement, since they do not matter at the moment.

This explication task also subsumes the grounding of indexical expressions - such as the deictic temporal or locative examples discussed above in Section 4.1.3. This grounding becomes necessary, for example, when inserting a real instance of a location in the *pragmatic gap* left by the omission of a starting position in Example 28. As many commercial mobile devices today, also the research prototypes employed in this work, use a global positioning system to supply the current location of the user which is observed and classified by the context model. Again, it is important to note that this type of situation-awareness is a necessary prerequisite for context-dependent analysis and that, for example, the task of determining the right level of granularity in classifying the position of the user is not a trivial one and context-dependent itself.

As a starting point, I will sketch out how this explication is realized in the multi-domain setting of the SmartKom system, where frame semantic collection and ontological modeling of the domains enable corresponding insertions directly on the domain objects models [Porzel et al., 2006a]. These individual domain-specific *common ground* models provide knowledge, which prevents, for example, a human travel agent from checking room vacancies without knowing the specific *instances of* at least an arrival date and the intended duration. This holds across several domains, as also a theater agent cannot reserve tickets without knowing the user-specific price range, location and date. In the following example - where

```
<informationSearchProcess>
  <entertainment>
    <performance>
      <cinema>
        <contact>
          <address>
            <town>
              here
            </town>
          </address>
        </contact>
      </cinema>
      <time>
        <beginTime>
          <at>
            now
          </at>
        </beginTime>
      </time>
    </performance>
  </entertainment>
</informationSearchProcess>
```

```
<contact>
  <x> 70.345 < /x>
  <y> 49.822 < /y>
  <town>
    Heidelberg
  </town>
</contact>
<time>
  <at> 19:00:00T26:08:03 </at>
</time>
```

Table 4.1: Context-specific insertions into a sample intention hypothesis resulting from the interpretation of a speech recognition hypothesis

the user is situated in a specific time and place, e.g. walking through the city of Heidelberg and initiates the exchange given Example 29, additional turns, e.g. asking the user to specify time and place, are avoided by decontextualizing the question and providing the answer given in Example 30.

(29) Was läuft im Kino
What is showing in the cinema

(30) Hier sehen Sie was heute in den Heidelberger Kinos läuft
Here see you what is running in Cinemas of Heidelberg

As specified above, the context model implemented in the SmartKom system enables the system to provide - hitherto implicit - knowledge concerning what is talked about.[12] The simplified structures given in Table 4.1.4 show insertions - given in bold face - into an SmartKom *intention hypothesis* made by the context model in the case of a question such as given in Example 29. In this case the explications inserted by the context model are threefold. Firstly, the cardinally required *place* node, corresponding to the range of a semantic relation such as *has-place*, was inserted into *cinema* object - as was the *time* inserted into

[12]Please note that, as discussed above, discourse context can - and does - override these structures, whenever prior utterances had already specified this type of information [Alexandersson et al., 2006].

the *performance* object. Secondly, corresponding object-specific instances are given as indexical *defaults*. These indexical placeholders are, then, contextually resolved. Here, the topical resolution of instance labels *here* and *now* enable the system to produce a suitable response, e.g. to display a map of the cinemas of an actual instance of a place, such as Heidelberg, and to present the specific performances of that day excluding those for which it is too late. The used instance labels, such as *here* and *today*, are replaced by means of function calls with actual values supplied by the respective external data providers that specify local position system and time. Last but not least, the domain-specific models specify how the received information is to be classified, i.e. how to resolve *here* with an appropriate level of geographic granularity, e.g. on a country-, town- or street-level, in much the same way as *today* is also replaced with a granularity-specific temporal *region*, such as a specific date and time or even just a season, as I will exemplify in Section 4.3.

By means of explicating such information and providing topical and contextually adequate values, the system as a whole is enabled to retrieve appropriate information from web sites or databases on what is currently playing in town or to produce maps featuring cinema locations, without asking the user to specify time and place. Instead the requested information can be given directly. Moreover, it also becomes possible to offer further assistance in making seat reservation or getting there. As mentioned throughout this work, the domain of spatial data, information and knowledge, especially as it applies to providing spatial instructions and spatial descriptions, constitutes an integral domain for the functionality of mobile dialog systems.

Concluding this initial discussion on the challenge of modeling and explicating implicit information faced in understanding conversational speech, I will now turn to the task of evaluating competing semantic specifications or *intention-hypotheses*, using empirically derived models of the pertinent pragmatic knowledge. This experimental setting is situated in the spatial domain and will be introduced together with its empirical data and performance results in the following Section 4.2.

4.1.5 Roadmap

I have presented a discussion of interlocutionary and situational effects in conversational dialogs and exemplified the problem of underspecifications in computer-directed speech given the tasks and domains of the mobile research prototypes employed. In Chapter 3 I presented a set of experiments on speech recognition noise and semantic ambiguities in which this contextual computing approach relied on the contextually appropriate domain- and discourse models. In the following, I will present a corresponding empirical examination of the problem of pragmatic ambiguity, whereby, I will again seek to elicit the contribution of including contextual factors in this approach, only this time these factors relate to properties of users and situations, i.e. they become pertinent as the interlocutors of these systems are situated in a given real-world context. After this empirical examination given in Section 4.2, I will assemble the individual pieces

introduced herein into a model of pragmatic patterns in Section 4.3.

4.2 Using Situational Context for Underspecification

As stated before, usually unaware of the potential ambiguities inherent in their utterances, speakers let the shared context bear the burden of disambiguating and fully specifying their messages. Furthermore, they trust in the addressee's ability to extract that meaning from the utterance that they wanted to convey. Therefore, in order to interpret the utterance correctly, the addressee must consider pertinent contextual factors. The speakers, again, anticipated this and constructed the utterance knowing - quasi habitually - that certain underspecifications are possible since the hearer can infer the missing information and that ambiguities are resolved by recourse to the shared common ground [Krauss, 1987].

In the following I will present findings from experiments tailored towards identifying and learning contextual factors that are pertinent to understanding a user's utterance in a mobile conversational dialogue system. Such systems frequently supply touristic and spatial information [Malaka and Porzel, 2000, Johnston et al., 2002, Wahlster, 2003, Ankolekar et al., 2006]. In the data collected in the course of the corresponding research [Porzel and Gurevych, 2002, Schiel et al., 2002, Kaiser et al., 2006, Mögele et al., 2006] one can find instances of phenomena labeled as *pragmatic ambiguity*.

In a way these examples constitute *bona fide* cases for contextual interpretation after phonological and semantic processing has been concluded. Before showing how further natural language analysis can incorporate specific situational factors, based on models derived from the empirically collected data, I will provide the standard linguistic differentiations made for the showcase ambiguity examined below. The aim, again, is to enable a context-dependent analysis of the given utterances in such as way that it increases the conversational capabilities of dialogue systems, by letting them respond in a felicitous manner more frequently than without this type of contextual computing. Considering the two different kind of responses given in Examples 31 and 32, one finds that the first constitutes a spatial instruction, while the latter is considered a spatial description [Klabunde et al., 1999].

(31) In order to get to the cinema you have to turn right and follow the Hauptstrasse

(32) The Cinema Gloria is near the marketplace on the Hauptstrasse

A spatial instruction as in Example 31 instructs the interlocutor how to get from one location/object to a different location along a specific path, which can be for example the shortest, nicest or fastest possible. A spatial instruction is a felicitous response to a corresponding *instructional* request. Furthermore,

natural language generation and spatial planning systems exist that can produce incremental spatial instructions for mobile users [Kray and Porzel, 2000, Jöest et al., 2005]. A spatial description, as in Example 32, tells the interlocutor where an entity to be localized is situated with respect to one more reference entities via appropriate spatial relations. A spatial description constitutes an appropriate response to a corresponding *descriptive* request. For natural language generation systems localizations can be formulated for various object configurations ranging from small scale objects [Wazinski, 1992] to environmental ones, where also contextual factors, stemming from the user and the situation to be described are considered [Peters, 1993, Klabunde and Porzel, 1998].

One can, therefore, say that a spatial instruction is an appropriate response to an instructional request and a spatial description, e.g. a localization, constitutes an appropriate response to a descriptive request. Responding with one to the other does not constitute a felicitous response, but can be deemed a misconstrual of the questioner's intention or an *intention misrecognition*. In all dialogue systems cases of intention misrecognition decrease the overall evaluation scores, since they harm the dialogue efficiency metrics, as the user is required to paraphrase the question, resulting in additional dialogue turns. Furthermore, satisfaction measures decrease along with perceived task ease and expected system behavior.[13]

4.2.1 The Task: Pragmatic Disambiguation

Having introduced these different types of responses - localizations and instructions - I will now examine the type of question, presented in Example 33, to which they constitute possible answers.

(33) Where is the cinema Europa

As one can see by looking at the empirical data - presented below in Section 4.2.2 - in real situations seemingly *simple* questions, such the one given in Example 33, cannot be understood by means of always construing them in one specific way. To do so - when there are at least two different observable ways of responding to it - means misconstruing either all instructional requests as descriptive ones or *vice versa*. As it is possible to misconstrue the intended meaning of a lexical item, e.g. whose meaning has been coerced [Hobbs et al., 1990, Michaelis, 2001],[14] one can misconstrue the overall intention

[13] Unfortunately, in the PARADISE dialogue quality metrics are not directly effected by intention misrecognitions, as they are not counted as such [Walker et al., 2000].

[14] A set of lexical phenomenon, that have correspondingly been discussed in Section 2.7.2, occur do some degree in the data stemming from more conversational systems. They concern, by and large, lexical ambiguities, as examined in Section 3.4. Other forms of coercions, such as in bridging expressions are found infrequently, while metonymies and metaphors are not found in the data examined herein. Nevertheless, both theoretical and computational work on texts featuring these phenomena exist [Lakoff and Johnson, 1980, Hobbs et al., 1988, Markert, 1999, Gedigian et al., 2006]. Among these *out of scope* phenomenon for the systems at hand, are also hypotheticals, counterfactuals and other mental spaces [Fauconnier, 1985, Faucounnier and Turner, 1998], even though computational progress therein has also been driven by constructional approaches to language understanding [Mok et al., 2004].

of the utterance, and thereby the *function* of the entire declarative or interrogative utterance at hand.

Therefore, to say that asking, for example, whether a *where interrogative* is construed as an instructional or a descriptive request corresponds to finding out if the questioner seeks to receive a spatial instruction or -localization as an answer. As I have stated above and will examine below, speakers habitually rely on shared contextual information and common pragmatic knowledge to be employed by their interlocutor to resolve the intention behind the produced speech-acts appropriately [Allen and Perrault, 1986, Perrault, 1989]. There are, again, two possible ways of approaching this task in natural language processing systems:

- One is to ignore it, viable for those cases where the application domain features a fixed set of singular construals, for example, when it suffices for a system to understand one type of request per interrogative. As a consequence systems become inherently single-domain and therefore unscalable in that respect. Also the input needs to be restricted to a less conversational one, e.g. by adopting controlled dialog strategies.

- The other is to attempt to resolve such *functional construals* appropriately, as in the contextual computing solution proposed below, thereby making systems more scalable and more capable of dealing more robustly with conversational language.

As in the classification tasks examined above, a set of attributes and values need to be defined in order to classify the individual markables. In this task the value for each utterance, serving as markables, is constituted by the construal of its pragmatic function, i.e. to determine what is requested in this situation. Attributes for such a value as *construal* can be found by means of firstly collecting and categorizing the responses one receives to these utterances in real situations. For example, if two types of responses, e.g. localizations and instructions, are given in response to the *where-interrogative* featured in Example 33, then the corresponding attributes for the value *construal* are *descriptive-request* and *instructive-request*. Once the attributes are given one can seek to find the best-fitting pragmatic construal given the observed context at hand.

4.2.2 The Data: Collection & Annotation

In the following I will present the empirical data and experiments conducted to collect situated utterances and determine the pertinent contextual factors that enabled the speaker to be pragmatically ambiguous and the addressee to construe the intended meaning correctly.

Collecting Questions

In an initial data collection American native speakers - who were abroad in Germany - were asked to imagine that they are tourists in Heidelberg equipped

Table 4.2: Instructional request types and occurrences in Corpus ASK_1

Type *Example*	#	%
(A) How interrogatives, e.g., *How do I get to the Fischergasse*	38	30%
(B) Where interrogatives, e.g., *Where is the Fischergasse*	37	29%
(C) What/which interrogatives, e.g., *What is the best way to the castle*	18	14%
(D) Imperatives, e.g., *Give me directions to the castle*	12	9.5%
(E) Declaratives, e.g., *I want to go to the castle*	12	9.5%
(F) Existential interrogatives, e.g., *Are there any toilets here*	8	6%
(G) Others, e.g., *I do not see any bus stops*	3	2%

with a small, personal computer device that understands them and can answer their questions. Among tasks from hotel, train and restaurant domains subjects also had to ask for directions to specific places. The resulting set of recorded and transcribed utterances, i.e. Corpus ASK_1, features 128 instances of instructional requests out of a total of approximately 500 requests from 49 subjects. The types and occurrences of the corresponding linguistic categories are presented in Table 4.2.

As can be seen from the almost equal distribution of *How-* and *Where-interrogatives* in the corpus, the subjects were as likely to express an instructional request using the ambiguous form of *Where-interrogatives*, which - this potential construal notwithstanding - still constitutes the most frequent form for expressing a localizational request. For example, numerous instances of *Where-interrogatives* requesting spatial localizations can also be found in other corpora such as the Map Task Corpus [Anderson et al., 1993].

The research prototypes that sought to handle more conversational speech and serve as experimental platforms for this work, specifically the DeepMap [Malaka and Zipf, 2000] and the SmartKom [Wahlster, 2003] systems, had implemented semantic parsers capable of interpreting some utterances as instructional request and others as descriptive ones. In the Deep Map system the utterances were analyzed using a semantic grammar [Gavalda, 1999] that produced individual speech acts by unifying typed feature structures [Carpenter, 1992], while SmartKom relied on a production system approach [Engel, 2002]. The respective grammars identified categories A, C, D and E as instructional request and *Where-interrogatives* of type B as localizational request. Obviously, since all *Where-interrogatives* look alike form wise, regardless of they where intended to be construed, extra-linguistic knowledge is needed to differentiate - or classify - them according to the actual request type at hand.

Taking the corpus of instructional requests presented in table 4.2, this state of the art results in recognizing roughly 63% of the instructional requests contained in the corpus as such. Changing the grammars to treat type B as instructional request would consequently raise the coverage to 92%. However, as stated above, *Where-interrogatives* do not only occur as requests for spatial instructions but also as requests for spatial descriptions, i.e. localizations of the named entity.

The problem faced is that, given only the linguistic forms without additional

context, the parser grammars can either interpret all *Where-interrogatives* as descriptive requests or as instructional requests. This implies that both systems can either misconstrue 29% of the instructional request from the corpus as descriptive requests or misconstrue all descriptive request as instructional ones. The corpus data therefore provides a performance-based baseline, as discussed in Section 2.6.2, of a precision of .63 in classifying instructional requests. Also, as a result of the experimental set-up, these utterances are predominantly discourse initial. This means that, even if discourse context could provide sufficient information to disambiguate such interrogatives correctly, it could only do against the backdrop of cohesive sequence of prior utterances and not when the user initiates the dialog by asking such a question.

Collecting Answers

As stated above, after defining the task-specific values and their attributes one needs to examine what factors might differentiate situations where one or the other construal ought to be favored. Or, to put it bluntly, but in the appropriate order, for science, having found and defined the problem, one can research possible solutions for it. As discussed above, additional information must be considered when one seeks for an appropriate construal of a given *Where-interrogative*. As the data discussed below show, an approach solely based on domain context alone, where primary distinctions regarding the object-type of the named entity are given reductionistically, e.g. whether the form is an instance of a building or a street, will not always suffice to solve the problem.

Given this task of finding out what could possibly matter for this question and other questions, an new experimental design was applied to gather empirical data in a situation that is as *unperturbed* by the experiment itself as possible. In this so-called *Field-Operative Test*, a set of operatives where hired, who went around in the city of Heidelberg an we asked people on the street specific questions, including the slightly more polite form of the *Where interrogative* as given in Example 34.

(34) Excuse me can you tell me where the Cinema Europe is

As any kind of audio- or video taping of the passerby's answers would have required their explicit consent, thereby changing the situation significantly, only contextual factors describing the operative's situation were manually logged together with a linguistic classification of the responses received by the operative. More specifically, the operatives kept track, as in a so-called *diary study*, of the following information, that can be inserted paradigmatically into variable [slots] of the sentence:

> On the specific [day], in the [time of day], I [operative], asked the [interrogative] with the [named entity] when this [proximate] to it.

Hereby, the slots where to be filled with the type of information presented in Table 4.3, which included also marking the gender of the person they asked and if

Table 4.3: Contextual Information about the Situation and Interlocutor

Situational Slots		Instance Fillers
day	←	the actual date of the experiment
time of day	←	morning, afternoon or evening
named entity	←	the castle, the train station, a specific hotel, a specific square, a specific cinema, an cash machine and a toilet
precipitation	←	rainy, overcast, sunny
temperature	←	colder (< 10 degrees), medium (10 - 20 degrees) and warmer (> 20 degrees Celsius)
accessibility	←	open, closed
Interlocutionary Slots		Instance Fillers
passerby gender	←	male, female
passerby age	←	younger (< 25 years), medium (25 - 50 years) and older (> 50 years of age)
operative id	←	the id of the field operative
operative proximity	←	near (< 5 minutes-), medium (5 - 30 minutes-) and far (> 30 minutes walking distance)
operative props	←	none, with bags, with bicycle,

that person seemed to be of younger, middle, or older age to them. Analogously, they recorded how the weather was at the time, in terms of precipitation and temperature judgments, if the named entity was open or closed and whether the they (the operatives themselves) were carrying bags, pushing a bicycle or not. Last, but not least, they noted what type of answer they received and - for some cases - indicated special features of the individual responses as free text.

In this experiment, the types of interrogatives were limited to types A, B and F, i.e., *How-*, *Where-* and *Existential-interrogatives* as exemplified in Table 4.2. Despite its low frequency in Corpus ASK_1, *Existential-interrogatives* were included, as questions of the type presented in Example 35, constitute a classic example of an indirect speech act [Searle, 1975].

(35) Gibt es hier eine Bäckerei
Is there a bakery here

Over the course of several weeks two operatives collected the contextual information presented in Table 4.3 and classified responds types to 364 questions posed. Together, 167 *How-interrogatives*, 128 *Where-interrogatives* and 69 *Existential-interrogatives* can be found in the resulting Corpus ASK_2. In terms of the text types received as a response to these questions, the operators classified 263 instructions, 79 localizations and 22 cases where the interlocutor gave both text types in one *hybrid* response.

Table 4.4: Types of Questions and Answers by Field Operative

Question	Type A (How)	Type B (Where)	Type F (Existential)
Op1	117	100	31
Op2	50	28	38
Answers	Instructions	Localizations	Hybrid
Op1	174	50	22
Op2	89	29	0

Table 4.5: Types of Answers by Questions

Question	Instructions	Localizations	Hybrid	Total
Type A (How)	165	0	2	167
Type B (Where)	60	54	14	128
Type F (Existential)	38	25	6	69

Learning What Matters

Unsurprisingly, instructions constitute the most frequent value in the corpus ASK_2, as there is hardly any other response to the most frequently asked *How-interrogatives*. From those 167 interrogatives asking explicitly for directions to places only four were not answered by instructions, but by means of the aforementioned hybrid form. A further examination unveils that these hybrid responses to *Where-interrogatives* are exclusive to situations where the goal object was a toilet. This was also the goal in most of the 25 cases where an *Existential-interrogative* was answered by means of a localization, which, given 6 hybrid responses, also means that - taking the remaining 38 cases - instructions were also the most frequent response to this *indirect* speech act. Furthermore, the corpus contains 128 responses to *Where-interrogatives*, which consisted of 60 instructions, 54 localizations and 14 hybrid forms. The operator-specific distribution of the questions posed and responses received is given in Table 4.4 and I also provide an overview of the response types gathered per question type in Table 4.5.

These data, therefore, support the previous finding regarding the pragmatic ambiguity of *Where-interrogatives* and other forms, such as *Existential-interrogatives*. Furthermore, they indicate that, in these situations, the employment of *How-interrogatives* was - on a functional level - practically unambiguous. A more fine-grained analysis shows that additional observations, made by the operatives, show differences in the *fleshing-out* of the particular instructions given, e.g. suggesting alternative transportation or questioning some aspect about the actions proposed. I will return to this type of underspecification in the following Section 4.3.

As stated above, this corpus of questions and answers together with the contextual information collected in the field contains 128 instances of the operatives asking the showcase *Where-interrogative* presented in Section 4.2.1. I have also

provided a discussion in Section 2.4.2 on the importance of determining what is context-variant and what is invariant for the given task at hand in natural language processing systems. This is especially, true when looking for pertinent contextual factors in the light of the richness of unrestricted language use in real situations. One feasible approach to finding potential factors that contribute to the underlying construal decisions made by the interlocutors - and can therefore be used to *predict* when it is contextually more appropriate to construe a given *Where-interrogative* as an instructional or locational request - is to perform an information theoretic examination of the gathered dialogical and contextual data [Porzel and Strube, 2002].

For this, each instance of the gathered data can be turned into a vector format, that represents the interlocutionary and situational context - using the attributes listed in Table 4.3 - that was observed when the interrogative was asked together with the received response, as the correct *answer*. Employing the entropy-based c4.5 machine learning algorithm [Winston, 1992], one can analyze which factors separate the different responses and to what degree they do so. Such an analysis, performed on the data gathered and described above, yields several noteworthy findings, that are *contained* in the decision trees produced by the learning algorithm, that relate individual contextual factors to the way the pragmatic ambiguity was resolved.

For example, next to the fact that the object type of a toilet seem be a kind of *outlaw*, which may be due to - still implicit - social factors, also permanently accessible places, such as public squares are not localized, but serve as the goal of spatial directions. However, objects that are currently closed, e.g. a cinema in the morning, are answered by means of localizations, whereby a few subjects explicitly asked the operative whether she actually wanted to go there at that time, and hardly by instructions. Nevertheless, if the object is currently open, e.g. a store or cash machine in the morning, people responded with instructions, unless the goal object is near and can be localized by means of a reference object that is within line of sight or when it happens to be in sight itself. In this case responses, as given in Example 36, were received, which - in terms of their text type - are localizations.

(36) Es ist da drüben
It is over there

Looking at the problem of finding an appropriate construal and, consequently, a felicitous response, for a given *Where interrogative*, one can see already that, depending on the combination of several contextual features, i.e. the object type, its accessibility and the given proximity to it, responses were either instructions, localizations or clarification questions. However, please note that by means of introducing additional contextual variations, e.g. having the operative dress as a craftsperson carrying buckets of paint, one could find more instructions to objects such as discotheques or cinemas even if they happen to be closed at present. This, as stated before, is a crucial challenge for contextual processing approaches, i.e.:

- to select the contextual data to monitor, as the *background* of the *frame*, e.g. the positions of the interlocutors or the situational accessibility of the individual named entities;

- to include them in the construal process, of what is in the *foreground* of the *frame*, e.g. a given utterance.

As noted throughout this work, what matters contextually for a given dialogical situation depends greatly on the specific task at hand, which, in turn, defines the domains about which specific contextual information and pragmatic knowledge is needed. This question will also become pertinent again when seeking for a way of describing situations formally, which I will discuss in Section 4.3. Before doing so, I will conclude this section by describing how such dynamic contextual observations and their consequences have been implemented in the SmartKom framework and what the ensuing results of applying this contextual approach are for scoring alternatives construals of the showcase ambiguity discussed above.

4.2.3 The Algorithm: Scoring Construals

Prior to casting these contextual considerations into the light of formal knowledge representation, I will present how employ graphical models can be employed to relate the contextual observations made in the experiment described above to the construal decision at hand. In this way contextual analysis can be performed by means of including correlations and their strength observed in the gathered data as conditional probabilities in a so-called *Belief-* or *Bayesian* networks employing a generalized version of the variable elimination algorithm [Cozman, 2000, Bryant et al., 2001]. Together the model's nodes, arcs and conditional probabilities represent the factors and their empirically observed relations to the decision at hand. Thus, they can be employed to compute the posterior probabilities of the decision at hand.

Generally speaking, graphical models are well-suited for combining heterogeneous, independent and competing input to produce discrete decisions and are used as mathematical abstractions to model specific cognitive processes underlying the way speakers process natural language [Narayanan and Jurafsky, 1998]. The simplest network possible estimating the likelihood of whether a given *Where interrogative* is preferably construed as an instructional or descriptive request, consists of three *observation nodes*. These nodes *observe* whether a *Where interrogative* is at hand, the goal object is open or closed and its proximity to the user. Additionally, there is one *decision node* connected to each of the *observation nodes* - when queried this node *decides* whether a spatial localizations or spatial instructions constitute a better fitting response.

As in the case of the multimodal system employed herein, interfaces to external sensors and services provide that contextual information. For example, within the DeepMap system [Malaka and Zipf, 2000], several relational databases, e.g. the *Tourist-Heidelberg-Content Base* and a geographic information system supply information about individual objects including their opening

and closing times. By default, objects with no opening times, e.g. streets and squares, can be considered always to be open, unless made inaccessible by constructions or other circumstances. Additionally, a positioning system supplies the current location of the user, which is handed to the geographic information system that can computes the respective distances and routes to the specific objects. This, again, is an opportunity to note that this type of context monitoring is a necessary prerequisite for context-dependent analysis, as this enables contextual computing approaches to make dynamic observations of the factors determined as pertinent by the empirical data collected.

As stated before, these observations, captured by the monitoring modules need to be converted into a contextually adequate representation. For graphical models, corresponding nodes, such as *accessibility*, and their attributes, such as *open* or *closed*, can be specified together with their conditional probabilities using an extensible mark-up language, such as the Bayes Interchange Format [Cozman, 1998]. Together with a given representation of the utterance to be construed in some way, e.g. a parser's output, they constitute the input sources for the showcase model described herein. The resulting output constitutes a measurement of what the given utterance should be *construed as*, i.e. the contextual fit of the possible alternative construals. This list of ranked construals, e.g. a list of two decisions for a given *Where-interrogative* with their corresponding scores.

The individual *scores* represent the probability of the questions being construable as an instructive request given the evidence as well the probability of the questions being construable as an descriptive request given the evidence. This can then be employed to enable the system to respond with the contextually better fitting type of answer. For this, the original parser output has to be converted into the dialog system-specific representations of instructional or localizational requests.

4.2.4 The Results: Pragmatic Ambiguity

As I have shown in Section 4.2.2 for the case of Corpus ASK_1, the possible performance baseline systems, such as the aforementioned DeepMap and SmartKom system, misconstrue 37% of the instructional requests of this initial data set. More specifically, due to the way the specific grammars are defined, all requests of type B and E as presented in Table 4.2, will incorrectly be interpreted as localizational requests and type F is not recognized at all and instances thereof would cause the system to indicate non-understanding. In terms of performance measures, used throughout this work, this corresponds to a precision of $p = .63$ on corpus ASK_1.

Based on the interlocutionary and situational context data gathered as part of Corpus ASK_2, a context-sensitive approach, as discussed above, can enable natural language understanding systems to construe *Where interrogatives* to distal but accessible places as instructional requests. In the case of the original set of *Where interrogatives*, found in Corpus ASK_1, which had been uttered by native speakers instructed to ask for the way, this corresponds to a lowering of

the percentage of misconstruals to 8%. The corresponding performance gain of .19 or precision of $p = .82$ can be considered quite an improvement. Additionally, if one employs the data gathered for *Existential Interrogatives* in a similar fashion, there is room for an additional coverage of 6%, which would leave only 2% of the initial data set as unanalyzable for the system.

I stated that the implementation of such a model needs to represent and integrate the diverse information- and knowledge sources necessary for the type of context-dependent natural language analysis proposed herein. I have also exemplified - using the phenomenon of pragmatic ambiguity as a showcase problem - how a contextual computing approach can contribute to decreasing the amount of misconstruals or intention misrecognitions in conversational dialogue systems. Such an enhancement of the systems' performances hardly goes unnoticed in user satisfaction evaluations. While it is quite easy to imagine that having the misconstrual rate drop by 35% would be beneficial to several PARADISE criteria[15], it is unlikely that such a consistently *picture perfect* performance can be cashed out directly from the showcase system described herein. This can be the case for several reasons:

- additional contextual factors that were not considered herein, but are nonetheless pertinent;
- unforeseen trade-offs with other construal decisions, both in terms of generating false positives as well as true negatives in the corresponding classifications;
- additional construals of *Where interrogatives* that have not been recorded or are novel in that context.

These remaining grains of salt notwithstanding, as the approach described herein results in a ranked list of possible construals for a given utterance together with their estimated probabilities, one can define more than the two types of responses described above, even without changing the network itself. This can be achieved, for example, by introducing a minimum distance that the *winner* must have from the second in line, for cases where the posterior probabilities can be considered too close to each other. If, for example, the difference of the posterior probabilities of the *instruct - localize* decision is smaller than a given amount, the system could respond by asking the user a clarification question as provided in Example 37.

(37) Do you want to go there or know where it is located

As stated in Section 4.2.2, this is also a type of response found in the field operative tests. This, in turn, would also enhance the conversational capabilities of dialogue systems next to increasing their understanding capabilities and robustness. I have also shown that in cases of pragmatic ambiguity the whole utterance

[15]This can be the case for aspect of dialog quality, such as task ease and expected behavior as well as for dialogue metrics, due to a decrease in the number of turns necessary to achieve task completion.

form *looks alike* and that, therefore, additional information and corresponding knowledge is needed to enable dialogue systems to pick the most appropriate reading. I have also exemplified how this can be realized employing situational and interlocutionary factors learned from empirical contextual data of the situation and the interlocutors at hand. Such an approach is also congruent to parsing approaches that seek to resolve ambiguities that arise during semantic interpretation by allowing for radial categories as well as probabilistic analyses [Bryant, 2008]. This constructional approach also includes contextual considerations, focusing on discourse context as a *context builder*, e.g. space-building expressions [Fauconnier, 1985].

Discourse, naturally, constitutes the central context one needs to consider for evaluating the cognitively motivated system against empirical data stemming from reading time experiments. Within this general research framework, that is based on understanding human language before formalizing and implementing the corresponding computational systems at various levels of granularity [Feldman et al., 1996, Feldman, 2006], proposals have been made to include situation- and domain-specific information in the formalized knowledge needed for a *deeper* understanding that goes beyond *shallow* parsing approaches [Bryant et al., 2001, Porzel and Bryant, 2003].

In the final section to come, before concluding this work as a whole, I will re-cast the question of construing underspecified utterances - as it was posed in the beginning of this work[16] and examined in an empirically driven manner in the Sections above, in the light of seeking to formalize the knowledge needed for making dialog systems more conversational. I will, therefore, focus again on representational issues, starting - in a sense - where the discussion on domain models stopped in Chapter 3. An additional research challenge, that provides additional motivation therefore, stems from enlarging the scope of the system from a multi-domain to an open-domain scenario, as in the case of the aforementioned SmartWeb project [Wahlster, 2004].

4.3 Modeling What Matters

In this section two fundamental, but notoriously tricky, notions for multimodal dialog systems, such as presented in Section 2.4.2, will be re-examined as one of the central problems facing both applications in artificial intelligence as well as in natural language processing. These, often conflated, notions are those of context and pragmatics. Indeed, in many ways both notions are inseparable from each other if one defines pragmatics to be about the encoding and decoding of meaning, which, as shown in Sections 3 through 4.2, is frequently context-dependent on many levels of analysis.

This entails that pragmatic inferences made in the process of pragmatic interpretation, or pragmatic analysis [Bunt, 2000], are impossible without recourse

[16] Sections 1.1 and 2.7 contain the respective motivation and modeling challenge for which the specific natural language processing tasks are displayed in Table 3.1 in Section 3.1.5.

to contextual observations. The specific sources for these observations, in the terminology used and proposed herein, provide contextual information that is based on the data they monitor. The information alone - while necessary to have - does not express what this information means for understanding and responding to the given situated utterance that triggers the ensuing inferences. In more metaphorical words, any given *cake* of contextual information does not specify how it wants to be cut.

As I have stated above, the distinction between pragmatic knowledge, which is learned and stays relatively stable, and contextual information, which is observed and can change rapidly, is important for designing scalable context-adaptive systems, which seek to interact with human users and to collaborate intelligently with them. Having discussed and examined the role topical contextual information in the explication task presented in Section 4.1.4 as well as for the showcase of resolving pragmatic ambiguities in Section 4.2, I will now return to the modeling challenge of capturing the needed pragmatic knowledge of how these observed things matter. At the same time, the focus of the discussion will stay on the task of understanding conversational utterances, which become almost chronically underspecified in the open-domain scenario of the SmartWeb project.

As mentioned in the introduction, advances in mobile hardware, communication and sensing technologies and the evolving web technologies set the stage for entirely semantics-driven approach to connect mobile multimodal human-computer interfaces, such as presented above, to web-based services. Brought together in one ontological framework these interfaces and services provide conversational interfaces for interacting with and accessing semantically described information. Based on these developments the SmartWeb project seeks to realize part of the vision of ubiquitous interaction by laying the foundations for multimodal user interfaces to access distributed and composable Semantic Web services. In the following I will present and motivate the ontological choices made in constructing the representational backbone of the system, especially where they are motivated by the need for describing the pragmatic knowledge pertinent for situated questions about - more or less - anything.[17]

4.3.1 Ontological Choices & Patterns

Both for theoretical reasons, to be discussed below, as well as a response to issues regarding (re-)usability, interoperability and scalability, recent efforts in ontology engineering are based on clearly defined modeling principles, explicit ontological choices and the employment of modeling patterns.[18] One such prin-

[17] In this work, I will not describe on the statistical question answering pipeline that is employed in the absence of corresponding domain knowledge [Ankolekar et al., 2006], and only point out that the continuous addition of specific independent domain ontologies and the parallel work on extracting semantics from web information and textual data, support the view that, by making the appropriate ontological choices, scalability and portability of semantic technologies increases.

[18] This lead to corresponding recommendations by the world wide web consortium's *Semantic Web Best Practice Group*, specifically the *Ontology Engineering and Patterns Task Force*

ciple choice concerns the basic nature of the ontology to be developed, i.e. whether it is descriptive or revisionary. While revisionary ontologies attempt to develop models of the world-as-is, descriptive ones intend to model the world-as-perceived by embodied humans. Since the latter consequently look at human language and cognition, they are populated by ontological categories that are independent from evidences stemming from other areas such as physics and astronomy.

A revisionary ontology, however, by and large ignores linguistic and cognitive aspects to avoid ontological assumptions that would be considered debatable on scientific grounds. To provide a borrowed example [Cimiano et al., 2004], that looks the common notion to make a *descriptive* differentiation between things that are in space and events that happen over time. In a revisionary setting time can be only another dimension for objects, e.g. based on relativity theory. Consequently, the common sense distinction between things that are and things that happen should be abandoned for a view according to which everything extends in space and time.

Another central ontological commitment, is to make the ontology reductionistic or multiplicative [Masolo et al., 2003]. This question in a sense extends the Type-Role distinction discussed in Section 3.1.4. The choice hereby is either to allow for an entity in space and/or time to be more than one type of thing, i.e. to be multiplicative, or to disallow it and to be reductionistic. In the reductionistic case an entity can, for example, either be a building or a hotel, but not both, while multiple *isa* relations to both building and hotel are possible in the multiplicative case. I will motivate the choices made in the SmartWeb framework below, after pointing out that it constitutes a descriptive and reductionistic ontology.

Even more recently the notion of ontological design patterns has quickly become a central issue in ontology engineering research. In its original form the first type of pattern, so called *logical patterns* specifies ways of solving standard ontology modeling problems, such as how to model n-ary relations or how to employ subsumption makros [Gangemi, 2005]. The second type, so called *content patterns* feature applications of logical patterns, and as an instance thereof is composed of logical patterns and combinations thereof. Content patterns are concerned with specifying ways of representing everything that is not given by the logical vocabulary itself, while, for example, *isa* relations come with the logical inventory, *part-of* relations do not and are, consequently, part of the domain-specific content of the ontology.

4.3.2 Foundational and Ground Knowledge

Another central aspect in ontology engineering, for which I presented an initial exemplification in Section 3.1.3, is the choice of a *foundational* layer, which is primarily used to guarantee harmonious alignment of various independently

[W3C-OEP, 2005] as well as by other international ontological standardization efforts, e.g. with the WonderWeb project [WonderWeb, 2003].

crafted domain ontologies and, moreover, to enable future re-usability. This foundational layer provides the basic ontological distinctions, axiomatizations and design patterns for the development of further domain-independent and domain-specific layers of *ground* ontologies as well as additional layers of *descriptive* ontologies, which I will discuss in Section 4.3.3 below. This important distinction is primarily motivated as a corresponding ontological separation enables an ontology engineer to express *reified* contexts [Rast, 2007] at the level of concepts or relations.

That means one can employ the same modeling instruments, including logical- and content patterns, for describing entities as one employs for modeling ground entities. This, in turn, circumvents the need to resort to other logical instruments for describing these entities, such as to formulate so-called *theories* about the ground model. As exemplified in Section 2.2, including theories about possible worlds or - in a weakened form - modal propositions about possible *situations* [Barwise and Perry, 1983], requires universal algebra to express the semantics of the logical forms [Whitehead, 1898].

As stated above, the alternative approach, pursued herein, provides to possibility to employ the logical-patterns, i.e. the specific logical vocabulary, of the given foundational ontology, in the same way one does for the ground part. This approach, consequently, requires the employment of a foundational layer for linking the ground and descriptive branches of the integrated ontology, as depicted for the solution taken in the SmartWeb project in Figure 4.2. In this case the foundational layer of the SmartWeb Integrated Ontology (SWIntO) is based on the highly axiomatized Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [Masolo et al., 2002, Gangemi et al., 2002]

DOLCE features various pattern-based extensions called *modules*, e.g. an ontology of plans, an ontology of information object as well as module for expressing reified contexts - or descriptions - called *Descriptions and Situations* [Gangemi and Mika, 2003], which will be described more closely in the following section. Additional to this foundational and descriptive layer, a ground domain-independent layer is included which consists of a range of branches from the less axiomatic Suggested Upper Merged Ontology (SUMO) [Niles and Pease, 2001], which is known for its intuitive and comprehensible structure. Additionally, the SmartWeb integrated ontology features several ground domain ontologies, i.e. a SportEvent-, a Navigation-, a WebCam-, a Media-, and a Discourse-Ontology [Oberle et al., 2006].

4.3.3 Logical- and Content Patterns

As discussed above, the DOCLE module *Descriptions and Situations* provides a logical- and content pattern for representing reified contexts and states of affairs [Gangemi and Mika, 2003]. In contrast to ground entities, such as physical objects or events, the extension of a descriptive ontology to include different conceptualizations of these entities poses a challenge to the ontology engineer. The reason for this circumstance is the fact these that conceptualizations are taken to assume meaning only in combination with some other entity. Accord-

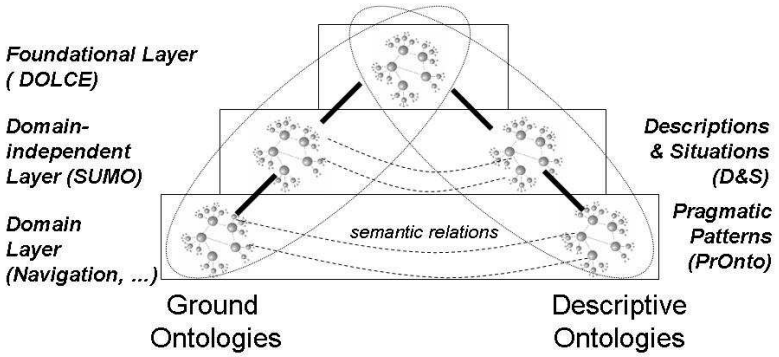

Figure 4.2: Foundational, Ground and Descriptive Ontological Layers

ingly, as discussed above, their logical representation is generally set at the level of theories or models and not at the level of concepts or relations.

In order to avoid potential terminological conflations and express is formally a descriptive statement is about something, e.g. it represents the *meaning* of some thing in a given context, while a ground statement is about the thing itself, e.g. for classifying instances within a given domain. In this modeling framework a situation is, consequently, clearly defined as a set of instances: For example, a situation could be constituted by the instances of a specific person, e.g. Rainer Malaka, a specific motorcycle, e.g. his African Twin, and a specific country road, e.g. the B3 in Germany, at a specific time, e.g. a certain day when it was sunny and 22 degrees Celsius. When seeking to describe this situation one would somehow like to express that Rainer was a motorcyclist, the motorcycle was the means of locomotion and the road was the path that he took on a nice day. In some other context however, e.g. that of an accident - where one might seek to describe him as the victim and the motorcycle as the means of injury - or in an environmental description - where he might be the culprit changing the weather, as the victim, by means of burning gasoline while driving on the B3.

In any case one would seek to refrain from simply multiplying the ground *isa* relations to express that, next to being a *person*, he is also a *motorcyclist*, *victim*, *culprit* and so on. Alternatively, employing the *type role* distinction applied in Section 3.1.3 leads to an explosion of roles for each domain-specific type that

would range over potentially all domain ontologies in a truly conversational open domain system. This approach, next to the fact that the *roles* proposed originally [Russell and Norvig, 1995] where not intended to serve as an instrument for describing the various context-dependent construals or conceptualizations of one entity that all hold true for it at the same time.

Next to avoiding these issues, employing a dedicated descriptive pattern for a context-dependent reification of ground entities, that is based on the foundational logical patterns for the contextual computing approach pursued herein yields additional engineering advantages. For example, the employment of design patterns, even by means of *refactoring* existing ontologies into pattern-based ones, has been shown to be beneficial for ontology quality when measured in terms of performance on a given task, e.g. ontology alignment [Svb-Zamazal et al., 2008]. Most importantly, however, a descriptive pattern for context-dependent reification, i.e. a coding of the functional meaning of some thing expressed in the dialogical terminology, introduced above, is to represent a pragmatically analyzed situation. It, therefore, enables the ontology engineer to express that, using the example provided above, someone is playing the *functional role* of a motorcyclist, who driving on a country road, that plays the role of the path on a day where the actual weather featured values that made it relatively nice.[19]

In the following I will consequently introduce the logical- and content patterns employed for capturing the pragmatic knowledge describing, for example, the functional roles that ground entities can play in given construed course of events which features contextually pertinent parameters whose values are part of the ground model in much the same way as courses of events are linked to ground processes and functional roles to objects. Before this, I will point out that - in its motivation - this approach to model non-physical conceptualizations not at the level of theories or models but at the level of concepts or relations [Gangemi and Mika, 2003], i.e. in the same way as physical objects and other first-order entities, is also based the neurological insights concerning embodiment and cognitive linguistic insights showing how we *manipulate* conceptual entities, as mental models, quite similarly to the way we manipulate physical entities as embodied beings. This, in turn, becomes relevant when seeking to *flesh out* what matters as discussed in Section 2.2. Pragmatically speaking, also in many cases even the relations and axioms modeled and applied for ground physical entities are also valid for non-physical context-dependent conceptualizations thereof. As shown in linguistic analyses of metaphorical and metonymical mappings between either the physical source and conceptualized target domain or different mental spaces [Lakoff and Johnson, 1980, Fauconnier, 1985, Mok et al., 2004].

[19] As is it bad practice to employ the term *relative* without specifying what it is relative to, let me add that it is precisely what is being captured by the over all pattern, i.e. that some specific observed values can mean good weather for motorcycling or bad weather in some other context.

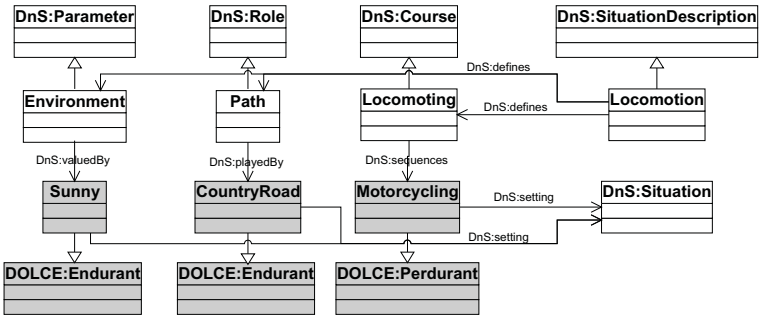

Figure 4.3: Model for the description "Locomoting"

The Description and Situation Patterns

As stated and motivated above two additional modules of DOLCE were added to the integrated SmartWeb ontology, i.e. the *Descriptions and Situations* module and the *Ontology of Information Objects*, which both can be seen as part of a larger undertaking to model tasks [Guarino, 2006, Gangemi et al., 2004]. In the form employed herein the *Descriptions and Situations* framework provides an ontological model of reified contexts that comes in the form of an ontology module. As explained above, it can be considered an ontology design pattern for domain ontologies that require contextualization.

The specified logical- and content patterns of the *Descriptions and Situations* module feature three descriptive entities, i.e. the classes *Courses of Events*, *Functional Roles* and *Parameters* [Gangemi and Mika, 2003]. These classes are linked by means of relations, which specify that:

- *Parameters* are *requisite-for* their *functional roles* and *Courses of Events*;
- *Functional Roles* are the *modality-targets* in the conceptualized *Courses of Events*.

Finally, the classes can be linked to the ground entities they describe, via the following relations:

- *Courses of Events* are *sequenced-by Perdurants*, i.e. processes within the ground ontology, such as *Motorcycling*;
- *Functional Roles* are *played-by Endurants*, i.e. objects within the ground ontology, such as of type *Person*;
- *Parameters* are *valued-by Regions*, i.e. phenomena that are sensed on scales, such as *Temperatures*.

For endowing the ground ontologies with a pragmatic layer of context-specific descriptions, this elementary pattern was employed to construct an underlying

model of pragmatics patters as shown in Figure 4.3, describing how the ground entities partaking in a situation are reified in their specific contextualized scene. In order to further illustrate and show the model at work, I will present individual showcase applications thereof within the SmartWeb open domain dialog system.

4.4 Pragmatic Patterns

As stated above, especially for mobile multimodal dialogue system contextual information is of central importance as the user expects the offer of topical services, e.g. while navigating through a dynamically changing environment that features changing precipitation- and temperature levels and/or traffic- and road conditions. This alone makes the adequate inclusion of contextual factors intertwined with the corresponding pragmatic knowledge inevitable for the task of natural language understanding.

The necessity to couple extra-linguistic situative information with pragmatic knowledge in the domain of spatial navigation has been demonstrated above as well as in other domains, such as sports [Loos and Porzel, 2005]. In the spatial domain of instructions I showed how underspecified information, was systematically explicated and considered in responding to situated interrogatives as given in Examples 4 and 33. If, for instance, the contextualized question-answer data, gathered as described in Section 4.2.2, indicates that public means of transportation are included in the spatial instructions when it was raining and walking there when it was dry, the corresponding pragmatic knowledge should explicate that *walking* can play the functional role of *means of locomotion* as a modality target for a given course of event, such as getting to some place, when the weather permits, i.e. the requisite parameters describing the environment are given. Analogously, one can express that *public forms of transportation*, such as busses or trams, can play that role when it is raining.

As already exemplified above, in the case where the means of transportation is given, such as when someone on a motorcycle requests directions to some place, the same requisite environmental parameters can serve to explicate that in one context curvy country roads can play the functional role of a *path* in dry and warm conditions instead of straight highways in case of cold or wet situations. While these common sense examples may illustrate some striking cases, a closer examination of question-answer dialogs shows that virtually every utterance becomes underspecified in an open-domain context.

Looking, for example, at the questions shown in Examples 38 and 39, one can find that - taken out of context, e.g. without knowing the even the domain at hand, i.e. which type of sport or entity - is talked about, it is hardly possible to answer these questions directly.[20]

[20] In a side experiment students were asked a given set of completely underspecified questions, such as the ones. Despite the variety of answers received, there were some frequent common answers, for example almost all of the German students responded with *Schumi* to the question *Who is the fastest?*, referring to the current Formula I champion at the time.

(38) How often was Germany world champion?

(39) Where is the Albatross?

As I have noted throughout this work, such a problem can be handled by either restricting natural language understanding systems to a pre-specified domain and hard-coded mappings that reflect the implicitly assumed context of the system. One could also shifting the pragmatic explication and specification task back to the user, e.g. by requesting to further descriptions of what ground entity is referred to. This, again, produces less efficient, more cumbersome und not very conversational dialogs.

It is important, again, to keep in mind that while it is necessary for context-adaptive systems to access services that provide up-to-date weather and traffic information or to keep track of the current discourse domain, it is not sufficient to solve the problem. Returning to our example it would, of course, be trivial to configure a system such that a hard-coded mapping guarantees that the question's meaning is mapped to the domain of soccer or that *Albatross* must always refer to a bar in Berkeley.[21] However, such approaches neither scale nor are applicable for open-domain systems. The solution proposed in this approach to contextual computing is to providing an explicit descriptive model of capturing the pragmatic knowledge of what is talked about.

4.4.1 Implementing Pragmatic Patterns

One implementation choice that arises hereby concerns the question of how fine-grained such a description and relation hierarchy should be that links the corresponding descriptive entities, i.e. courses of events, roles and parameters, to elements of the ground ontology. Hereby, also the classic trade-off between modeling and axiomatization comes into play. In the latter case, i.e. if a corresponding axiomatization should bear the burden of differentiating the pragmatically described entities at their respective level of granularity, then a corresponding axiomatic pattern for any description of type *SituationDescription* (SD) via the predicate *is-pertinent* (isp) to the respective descriptive entities, as given in Axiom 4.1.

$$\begin{aligned} &\forall(x) \rightarrow SD(x) \rightarrow \\ &\quad isp(SD, Region) \wedge \\ &\quad isp(SD, FounctionalRole) \wedge \\ &\quad isp(SD, CourseOfEvents) \end{aligned} \tag{4.1}$$

Alternatively one can shift the burden to the modeling side an specify individual relations for each pragmatic description of a situation, which can be achieved by encoding the class label of the specific pragmatic pattern into the relations

[21] And even that *Berkeley* describes the town of Berkeley in California, if one wants to be explicate, what is left pragmatically explicit even further.

it features to the ground entities [Loos and Porzel, 2005]. In either case this application of the *Descriptions and Situations* module for explicating implicit information extends the notion of deriving an instance - in this case a situation - from a description by modeling a pattern of pragmatic knowledge.

For example, in describing the ways in which context of what is talked about can influence what constitutes an appropriate ground entity, for example for grounding the expression *world champion*. As all that is given so far is that it plays a functional role of a *rheme* in a corresponding utterance, where the *theme* is implicit. The linguistic notions of *Thema* and *Rhema*, employed here as defined by so-called *Prague School* of linguistics, which concerned itself with the communicative pragmatic structures of sentences [Ammann, 1928], correspond to the notions of *topic* and *comment* in examinations of information structure [Lambrecht, 1994]. Applying this approach to expressing what is, somehow, contextually given, as the theme, i.e. what is known - and expressing what is new as the rheme for reifying questions can be seen as an attempt to find descriptions for underspecified phrases. This effort has also becomes more feasible than before - not because language-use and situations did not happen together before - but since contextually pertinent information can be observed and classified as it happens together with a given question.

Again this classification can either be informed by empirically gathered and analyzed data, as shown in Section 4.2, or by expert knowledge and includes - next to situational and interlocutionary context includes also *observing* domain- and discourse context as discussed in Chapter 3. Together these types of contextual information can serve to situate a given utterance, such that is become a question of finding the appropriate descriptive pragmatic pattern. As for the question given in Example 38, if one regards the functions of Thema and Rhema as complementary, where one needs to find the theme, the implicitly known, given the new rheme that is in the foreground. Note, that no discourse element is *a priori* a theme or a rheme. Furthermore, this may change with every new utterance; i.e. rhemes usually become the themes of subsequent utterances. Speakers in an everyday situations can also employ context-dependent pragmatic knowledge about what's known and what's new, when asked to specify it in given utterances [Loetscher, 1984].

In our Example 38 the theme - i.e. what is already known - is the discourse domain, which could be the ground entity *Soccer* or any other sport. As it is not explicitly expressed in the utterance, it must be inferred by recourse to context; in this case a prime supplier of information is the discourse context, but also situational factors, as when the interlocutor location is in a specific sport facility. The rheme is *world champion*, which is the new (unknown) element of the utterance. Using the descriptive patterns one can now describe that when someone asks a question - for which a description modeling *Interlocating* as a course of event which sequenced by the ground real world process of *Questioning*. The descriptive entity that plays the functional role of the *Rheme*, in this case, is contextually dependent on the instantiation of the parameter *Theme*. In short, what the user actually wants know depends on the context of what is assumed

to be known.[22]

As this example shows, one can capture a seemingly simple piece of pragmatic knowledge, i.e. that what a person wants to know can depend on what is already known in a given situation, on a very general level. Moreover this model scales for both multi-domain or open-domain systems, as it allows any domain to value the parameter of *Theme*, any object to play the functional role of the *Rheme* in a course of events which sequences *Interlocating*, which then **describes** the process of asking a question.

4.4.2 Applying Pragmatic Patterns

As mentioned above this model of pragmatic knowledge and the corresponding component for context-dependent processing seek to enhance the conversational understanding capabilities of dialog systems. As question such as *How often was Germany world champion?* poses a challenge to conversational open-domain dialog systems, since the theme of the utterance is not made explicit by the utterance. As the application of the needed pragmatic knowledge should - in contrast to controlled dialog systems - enable the user to be able to make utterances in any domain of interest without placing the burden of explicating the pertinent contextual factors and of describing the situation on him or her. Therefore, a systematic and scalable way of using the pragmatic knowledge so that a correct or felicitous answer to such questions can be found and to do what any intelligent interlocutor has to do, keep track of the shared context and make the appropriate inferences.

In order integrate this knowledge with actual contextual observations, which as expressed in the second statement and can be regarded as an observational task assigned to the implemented context model. Nevertheless, as discussed in Chapter 2 most systems assume an implicitly given domain context or employ various shortcuts to deal with problems of underspecification. One reason is, simply, that takes quite an effort to keep track and make sense about what is happening and what is being talked about or, in our terminology, to observe the given ground instances of descriptive parameters, whether they be an utterance specific theme observed as discourse-specific information or the weather conditions as situational-specific information, which - as all contextual information - can change dynamically and even rapidly.

Fortuitously, in a mobile dialog system contextual information is of high significance for various reasons, as a user expects the offer of topical services, while navigating through a dynamically changing environment, that involved changing precipitation-, temperature or gasoline levels as well as traffic- and road conditions. This makes the adequate coupling of pragmatic patterns and contextual information both feasible - and, as discussed herein, necessary for

[22] Please note that this does not have to be discourse context alone, as such a question in a discourse initial position must be decontextualized by the other context types, e.g. both interlocutors are watching a soccer game on TV or - if no contextual clues are available - would, most likely, prompt a sort of *what do you mean?* reply unless some other evoked *default* context is chosen.

enabling context-sensitive processing that produces more conversational and felicitous interaction. The motivation for coupling contextual information with pragmatic knowledge in the domain of spatial navigation has been discussed above in Sections 4 and 4.2. Existing navigation ontologies contain ground route models [Malyankar, 1999], which do not capture contextual dependencies. Given a single application-specific context, e.g. guiding only pedestrians -always on foot and always on the shortest path, one can employ context-free ground ontologies. However, already if the system seeks to make use of the interfaces offered by today's route planning and navigational web services, one must provide the means to describe what the right settings are depending on the actual situation at hand. Moreover, one wishes to do so in the least invasive way, i.e. by minimizing the amount of parameters and specific settings that have to be obtained by bothering the user.

Obtaining Utterances in Context

In order to allow systems such as the SmartWeb prototype to employ a wide range of internal and external services and information providers a semantic layer was implemented that extends the ontology of information objects [Oberle et al., 2005], which itself is used to *express* descriptions by means of some form system. The context model used for observing contextual information is implemented as a module within SmartWeb's dialog manager. It interacts with the dialog manager's middle-ware. As stated above, in the entirely semantics-driven approach also the internal communication format in SmartWeb is based on RDFS, which I introduced in Section 3.1.1, and, therefore, consists of corresponding RDF instance-documents. The employed RDF schema was based on the world wide web consortium's proposed standard for representing multiple interpretations of multimodal input, it constitute, therefore, an ontological *version* of the extensible multimodal mark-up language (EMMA), which is called *SWEMMA* [EMMA, 2004, Reithinger et al., 2005].

In a SWEMMA document, as a collection of RDF instances, which are also called *triples*, the actual input and the possible interpretations are consequently represented as instances of the discourse- and a respective domain ontologies. Within the dialog manager these documents are stored as assertions. All dialog components can access these assertions via pointers to the root instance of an interpretation provided by the middleware. Each dialog component then adds its own interpretation to the EMMA document.

As described above, the context model receives the semantic interpretation via the middleware, after it has been processed by the modality specific recognizers, e.g. for speech and gesture, and their respective analyzers. The task for the model is to change the semantic representation in such way that context-sensitive explications are formally represented, as if the user would have done so explicitly.

As I have stated before in a mobile dialogue system contextual information is of central importance and makes the adequate inclusion of contextual factors to be *intertwined* with the corresponding pragmatic knowledge necessary for the

achieving the task of responding felicitously to the user's input in a way that scales to more than one type of situation. Before, showing how the modeled pragmatic patterns are *matched* against the observed contextual information, I will present the range of data and information available to the implemented context model.

Adding Context to the System

The interface to the sensor data encapsulates so-called *context sources*. These context sources are identified by respective hierarchically organized concepts from the ground ontology and provide the context information as instances of these concepts. In other words, this design pattern for representing contextual information provides a set of instances together their ground conceptualization as an *anchor*. Below, the set of sources encapsulated in this way are given:

- A global position system is connected to the user device and delivers current location data to the dialog manager, from where it is passed as a message to the context model in regular intervals. This spatial context source employs also an external web service to resolve the exact address using inverse geocoding.[23]
- A weather service context source encapsulates and polls a web service for current weather conditions depending on the current location or any other, e.g. as intended goal locations specified by the user.
- A traffic service encapsulates and queries a web service for current road conditions, depending on the source and goal locations specified by the user or the system.
- A time context source encapsulates time information from the real time clock
- A set of external sensors in vehicles, such as featured in the prototype car and motorcycle provided by the respective manufacturing project partners, are encapsulated as context sources. They observe information, for example, concerning gasoline levels, whether the tires are spinning or not as well as additional warnings.
- A discourse and domain context source provides pertinent information from current discourse- and domain context, as discussed in Section 4.1.3 in the light of the experiments presented in Chapter 3.

Again, let me note, that calling this input *information* implies that the *raw* data monitored has been cast into some - more or less - structured representation that classifies it. As in the case of the SmartWeb system, this frequently entails mapping an existing representational format, such as specified by the web service

[23] To avoid unnecessary network and computational load, this information can optionally be cached and updated only if the location has changed significantly.

description language [WSDL, 2001], into the corresponding system's ontological vocabulary [Oberle et al., 2005]. As in the case of ground entities of type *Region*, the ontological model can provide the corresponding classifications, which, in turn, then constitute the ontological ranges of the *valued-by* relations of the descriptive pragmatic patterns featuring the corresponding parameters.

As the information from the context sources are represented as classes from the ground ontology it provides instances of these classes or subclasses thereof. For example, the location context source specifies a *SpatialRegion* and delivers instances of *City* as a descendent of this ground region. I will, in the following, present, how these ground representations and descriptions are employed by the context model.

Adding Pragmatics to the System

For finding the appropriate pragmatic descriptions, the context model performs two passes over the instances contained in the SWEMMA documents found via the middleware, as discussed above. As these instances are part of the ground ontology and are logically connected to the respective descriptive entities via the aforementioned relations:

- *sequenced-by* that features the logical range of ground *Perdurants*, e.g. processes within the ground domain ontology, such as *Motorcycling*;
- *played-by* that features the logical range of ground *Endurants*, e.g. objects within the ground domain ontology such as *CountryRoad*;
- *valued-by* that features the logical range of ground *Regions*, e.g. phenomenon within the ground domain ontology such as *Temperature*.

As defined in the core design pattern, described in Section 4.3.3, these relations are associated to the respective descriptive entities featured in the pragmatic patters, which are of type *CourseOfEvent*, *FunctionalRole* or *Parameter*. These axiomatically or associatively connected entities provide a set P of situation descriptions, $SD_1 \dots SD_n$, as discussed and exemplified in Section 4.4.1. Consequently the entire set P, then, constitutes the descriptive pragmatic ontology, which I labeled *PrOnto* in Figure 4.2. As each descriptive entity E_i serves as the domain for the one of the set of relations listed above, it has - as its range - a specific ground entity G_i *connected* to it. As implemented herein, an individual typed ground entity G_i can *evoke* as many individual descriptions from the set P as the number of *connections* it features.

To give an example, supposing a ground entity G_x is modeled as the range of two *valued as* relations to two different descriptive entities E_x and E_y of type *Parameter*, which are contained in two pragmatic patterns SD_x and SD_n of type *SituationDescription*, then G_x can be said to *evoke* the pragmatic patterns SD_x and SD_n. The number of potentially pertinent pragmatic patterns, which consist of descriptive entities and their relations, therefore, can be defined as the number of times where the ground entity is found in at least one of the

ranges featured by one of the descriptive entities contained in the patterns. In other words, if the ground object does not serve as the range of any descriptive entity featured in a specific pragmatic pattern, it does not evoke that pattern, because it neither sequences a course of event, plays a functional role, or values a parameter in the context-specific model. If it does so, however, then it might indicate the context-specific type of situation at hand, because it could sequences a course of event, play a functional role, or value a parameter in it.

In the first pass, all correspondingly evoked descriptions are collected and put in an *active patterns* pool. If no pragmatic patterns are applicable the set of assertions that represent the multimodal input are not modified and the *message* is returned to sender via the middleware without any changes. As defined above, for a pragmatic description to be applicable means that any of the ground entities contained in the SWEMMA document has been connected to a descriptive entity of type *CourseOfEvent*, *FunctionalRole* or *Parameter* via the respective relations *sequenced-by*, *played by* or *valued by*.

This now sets the stage for the contextual processing operations needed for the last two of tasks that I presented and specified as contextual computing tasks in Section 4.1.4, and which are listed again below:

- selecting the *best-fitting* description
- explicating hitherto implicit information

In discussing these final tasks, I will re-cast, the problem of pragmatic under-specification and -ambiguity, in the light of the presented pragmatic model and its application in the SmartWeb project.

Finding the Best Descriptions

Returning one more time to the showcase of pragmatic ambiguities, as given in Example 33 and discussed in Section 4.2, it is now possible to express two distinct descriptions:

SD_i for describing an situation featuring a *InstructionalInterrogative* as a subclass of the more general description for *Interrogative* situations, as exemplified in Section 4.4, which itself is a subclass of the core descriptive entity *CourseOfEvents* described in Section 4.3.3.

SD_l for describing an situation featuring a *LocalizationalInterrogative*, which also constitutes a subclass of *Interrogative* and *CourseOfEvents*.

If in both descriptions the classes *InstructionalInterrogative* and *LocalizationalInterrogative* are the domain of a *sequenced-by* relation that features the ground entity of a *WhereInterrogative* from the corresponding discourse ontology as their range, both descriptions are consequently evoked as exemplified above. Unlike each other, however, the descriptions posited here feature distinct ranges for the pertinent classes of type *Parameters*. As types of *Parameters*, such as *Accessibility* or *Proximity* can feature different ranges in individual descriptions

- either as a result of the employed logical pattern exemplified in Axiom 4.1 or the corresponding content pattern discussed in Section 4.4.1 - they can be *valued-by* different ground regions. The individual ground entities of type *Region* that serve as individual ranges, can be determined and classified based on learning experiments, as in the case of finding the values for the attributes accessibility and location, as discussed in Section 4.2.2, or determined and classified based on domain expertise. Analogously, as I will discuss further below, they can feature classes of type *FunctionalRole* which describe the roles the modeled objects play and what ground objects they range over, e.g. *Cinema* or even *Toilet*. Moreover, they do so whether the entities are explicitly mentioned or not.

Having these two distinct descriptions in the pool of *active patterns*, in other word means a situationally observed ground entity that was classified as an instance of a *WhereInterrogative* by means of the modality-specific analyzers can be construed as one or the other, making it pragmatically ambiguous. The next logical step, then, is to check which type of *Interlocating* can be specialized given the individual range restrictions of the descriptive entities contained in them, against the contextual information observed and mapped unto the ground vocabulary as described above in this section. The general notion behind this is that the more specialized a potential active pattern can become - depending on what instances of ground entities can actually be observed as real *participants* of the situation, the more pertinent contextual evidence there is for the most specializable description being a *plausible* description of the situation.

In this implementation a sequential collection of all possible specifications for all still specifiable active patterns is performed, by selecting one participating ground entity after the other and checking it against the range of the yet unspecified descriptive entities in the active patterns pool. As a result an underspecified descriptive entity can, consequently, become specified or stay underspecified in the process. After all participants' restrictions have been checked, the remaining *underspecified* descriptive entities can be considered *unspecifiable* given the contextually observed information at hand and all specifiable ones have been specified.

While checking each participant's restrictions - albeit in a sequential manner - all active pragmatic patterns are taken out of the pool and returned to it afterwards. Each of the *competing* patterns thereby remains active as long as there are other pertinent participants to be checked.[24] If there are no more specifiable descriptive entities this then also means that all participants' restrictions have been checked against the active patterns, which include can also alternative specifications of the same pragmatic pattern. In all case a specific amount of each pattern's descriptive entities has been specified more or less specifically, which provides the context model the means of selecting the most specified one

[24] In order to determine this *pertinence* the graph is traversed inversely matching the given range restriction's type, i.e. its super-class against the ground branches of the respective domain ontologies to find potential fillers that are, then, included in the set of pertinent participants, which can lead in some cases to multiple ways of grounding a descriptive entity [Babitski, 2004].

- or the most *active* one as the best-fitting pragmatic pattern for describing the given instance of a situation at hand. This pattern can, then, be considered as the most plausible *SituationDescription* or the *winner* of the set P.

In several ways this procedure is analogous to the approach for scoring alternative semantic representations presented in Section 3.2.3, only this time one can speak of alternative *pragmatic representations*. Please also note that the semantic representations contain ground entities while their pragmatic counterparts contain descriptive entities. As there are also multiple ways of implementing the corresponding scoring approaches presented herein, there are also multiple alternative approaches feasible. One can, for example, adopt a semantic-density based approach [Bryant, 2003] for a corresponding measurement of *pragmatic density* or even re-use the graphical models described in Section 4.2.3, which can be achieved by means of linking ontologies to Bayesian networks [Ding, 2005]. Before returning the these alternative approaches and corresponding experimental examinations thereof, analogous to the ones performed for the algorithms employed herein, I will conclude the description of this implementation, by the last step in the contextual processing pipeline, i.e. the explication that ensues with selecting a most plausible pragmatic pattern.

Augmenting the Ground Representations

In this last step, triples representing the specified ranges of the other descriptive entities of the most plausible pattern, but not given in the corresponding ground structure already, will be added to the RDF instances of the SWEMMA document that evoked the winning pattern. This, again, is familiar, as I have exemplified the corresponding operations, which can also be performed on XML documents, in Section 4.1. When the context module can apply its pragmatic knowledge as described above, it will pragmatically enhance the given semantic representation of the user's utterance in a context-dependent manner. This is done by specializing a concept or inserting, hitherto implicit, instances into the structures representing an interpretation of the user's multimodal input.

Ultimately the included information arrives via the encapsulated connections to the individual context sources, that provide situational and interlocutionary information as well or information about the domain- and discourse-context at hand. An overview of the employed sources is given in Section 4.4.2.The module, therefore, communicated with web services for topical weather, road conditions, and the current position just as can communicates with other components of the system to obtain discourse-specific information or domain-specific semantic measurements on ground domain models as discussed in Chapter 3.[25] In this way the module can apply the pragmatic knowledge to enhance the semantic representation of given input.

[25]The module is, therefore, connected to other dialog processing modules, i.e. for speech interpretation, multimodal fusion and dialog management as well as the reaction and presentation manager and the EMMA Unpacker/Packer that handles communication with the other multimodal recognizers and the semantic mediator which manages access to specific knowledge services, within SmartWeb's multimodal dialog processing architecture [Reithinger et al., 2005, Porzel et al., 2006c].

On the implementation part, it can also be noted that the model applied above featuring various descriptive pragmatic patterns for different domains, e.g. navigation, discourse or sports. Each description may span over various domain ontologies, e.g.:

- employing the Navigation Ontology and Discourse Ontology to model pragmatic patterns involved in understanding spatial requests;
- employing the SportEvent- and Discourse Ontology to model the pragmatic patterns involved in talking about sports.

These models consequently find employment in understanding navigational request and context-dependent route planning as well as in understanding questions about sports.

In order to summarize this implementation the input can be listed as follows:

- the descriptive pattern-based pragmatic ontology;
- the ground results of the SmartWeb semantic multimodal analysis;
- the ground information provided by different context sources.

In those cases where ground objects evoke to pragmatic patterns the module uses topical information to perform a selection and corresponding decontextualization, i.e. it then returns the ground results of the SmartWeb semantic multimodal analysis augmented with pragmatically inferred ground entities from the pertinent context sources. In order to evaluate the contribution of this approach, I will discuss how this can be assessed in the following Section.

4.4.3 Experimental Results: Decontextualization

Evaluations of dynamic context-adaptive processing techniques pose similar challenges as other means of adaptiveness. This is due to the fact that, given constantly changing contextual conditions, it is difficult to obtain enough contextual snapshots of the world in which a situated dialog takes place and to craft a corresponding evaluation gold standard. Nevertheless, I have already presented evaluations of the contributions of including individual types of context into account in the respective approaches proposed and described above. That is in cases where it was feasible to craft independent experimental training and test data and to create respective gold standards.

Assessing the potential contribution of the decontextualization procedure outlined above, however, is also not a matter of adding the individual contributions observed before. Ultimately one would seek to test such a system in real situations using the final SmartWeb mobile prototypes, which unlike the SmartKom system was not examined in an end-to-end evaluation, and only tested for qualitative usability in laboratory conditions throughout the

project.[26] As an alternative procedure to asses the potential contribution in some metric form one can ask the question if and how context would have mattered in a corpus of questions posed to the system. For cases where the *right* answers are only retrievable by recourse to contextual information, one can - at least - estimate the effect of the corresponding decontextualizations performed by the context module on dialog efficiency [Walker et al., 2000], by means of annotation experiments, as I will discuss in the following.

Data Collection & Annotation

For estimating how decontextualization can improve the performance of a dialog system on various user tasks, a corpus of typed queries was gathered by providing a web-based interface for people to type in questions they would pose to the system. The corpus contains 50 questions about soccer, soccer teams, matches or players, about 10 requests for route directions, cinema and weather information and 10 looking for pictures of a point of interest which the SmartWeb system can answer by recourse to a web cam service. The remaining questions are from other domains, e.g. as shown in Example 40. In total the Corpus ASK_3 consists of 100 questions.

(40) who invented the radio

As before, different annotators classified the queries as markable whereby for each context type the attribute *Specification* had to be annotated with the values *explicit*, *underspecified* or *irrelevant*. A context type was to be considered as *irrelevant* if the annotator cannot find a contextual setting in the respective type in which this sentence would have a different meaning. For example in annotating the sentence, given in Example 41 annotators found that situational context matters - that, depending on the location, different responses would be expected - while they also assign the value *irrelevant* to the domain context - that the answer should not change if talking about sports or entertainment.

(41) which direction to Berlin

Based on this classification one can estimate the number of insertions and additional turns a dialog system would need to decontextualize a question sufficiently as discussed below.

Experimental Results

I present the corresponding results of this annotation for the Corpus ASK_3 in Table 4.6 . All numbers are absolute values for the 100 markables contained in the corpus. As one can see, the domain is always either implicit or not relevant, which indicates the importance of recognizing and monitoring the domain context [Rüggenmann and Gurevych, 2004b]. Location context, on the other hand,

[26] This unfortunate state was caused by a combination of the high costs of running the experiments and the limits of the funding obtained for the project.

Table 4.6: Annotation of underspecified descriptive entities in the corpus

Context Type	Context matters	Context is underspecified
Domain	78	77
Time	86	47
Location	88	63
Weather	4	4

is in most cases not specified but, as expected, crucial for answering navigation requests. At the same time, in this corpus, actual speech-time was neither relevant nor explicitly given. Also for the open domain questions, the location context is almost in every case not relevant.

In order to test inter-annotator reliability, the data was annotated twice, achieving an absolute agreement - where the two annotators agreed on all values for the attributes of a given markable of 79%. This agreement metric means that if the annotators disagreed on a single value on any of the attributes annotated, it was counted as disagreement on the markable. The resulting kappa statistic [Carletta, 1996], is $\kappa = .72$ which can be considered as being quite reliable.

Using this findings one can estimate the bounds for improvement in terms of dialogical efficiency. In this case to request the missing information from the user, the task completion on our corpus would require 191 additional controlled dialog turns. Therefore, we can say that in this case the potential gain in efficiency lies at 48%. In other words a context-insensitive system controlled dialog system would require up to 1.9 more turns in average.

Additionally, in evaluating this approach one can examine the computational cost that comes with such a gain, as the context module itself consists of several main and auxiliary components, that manipulate ontological structures and perform graph searches [Cormen et al., 1990] next to observing the context sources. Using a cache for the observed context information ensures minimal latency for the dialog manager. For associating each instance in the interpretation of the user's utterance to the descriptive entities from the pragmatic ontology, the computational cost of the graph search can be minimized if all pragmatic relations are modeled in a bi-directional manner. As the concepts representing a description are linked by relations or axioms, the whole description can, then, be extracted and put in an active descriptions pool at practically no cost. Whenever a concept contained in a description ranges over one provided by a context source, the source will be queried, which can lead to various response times, depending on internal- or external web operations, which can be addressed by adaptive caching and pre-fetching techniques. The last iteration over the interpretations, where the instances are modified corresponding to most specific pragmatic pattern, is also computationally insignificant.

In the following conclusion, I will summarize the individual and aggregate

contributions made in this work in the light of the aims that motivated it. This overview will be followed by a discussion of future and ongoing work, which is concluded by final remarks on what has been made evident through this work.

Chapter 5

Conclusion

Firstly, I will summarize the results of this work in the light of the specified aims and intended contributions, as presented in Section 1. Therein, I stated that the central aim of this work is: "to present a formal approach for explicating contextual information and pragmatic knowledge that can be applied, employed and evaluated in natural language understanding systems".

Therefore, I specified and described *formal* and *explicit* knowledge models as machine-readable and logic-based conceptual specifications of a domain of interest [Gruber, 1993] in Section 3.1.1. Thereafter, I have presented contextual information as instances of observed data that is formally classified in ground logic-based representations, e.g. in the domain ontologies described in Sections 3.1.4 [Gurevych et al., 2006] and 4.3.1 [Oberle et al., 2007]. Lastly, I have described a model of pragmatic knowledge - employing pragmatic patterns, presented in Section 4.4, and linked them using a logic pattern that connects them with ground entities as discussed in Section 4.3.3 [Gangemi and Mika, 2003]. I have also showcased how empirical experiments can contribute to the process of determining how to link the two in Section 4.2.2 and how the combined model can be applied computationally in dialog systems as well as other natural language processing tasks [Porzel et al., 2006a].

In the course of evaluating and assessing the contributions - enabled via the respective ground domain models, the contextual information observed and the pragmatic knowledge modeled and their application - several ancillary empirical results were gained via the experiments performed for the individual and combined context types in:

- Chapter 3 where I presented experimental results of applying domain- and discourse context for a set of tasks in natural language processing on corpora SRH_1, WSD_1 and REL_1;

- Chapter 4 where I presented experimental results of applying user- and situational context for a set of tasks in natural language processing on corpora ASK_1, ASK_2 and ASK_3.

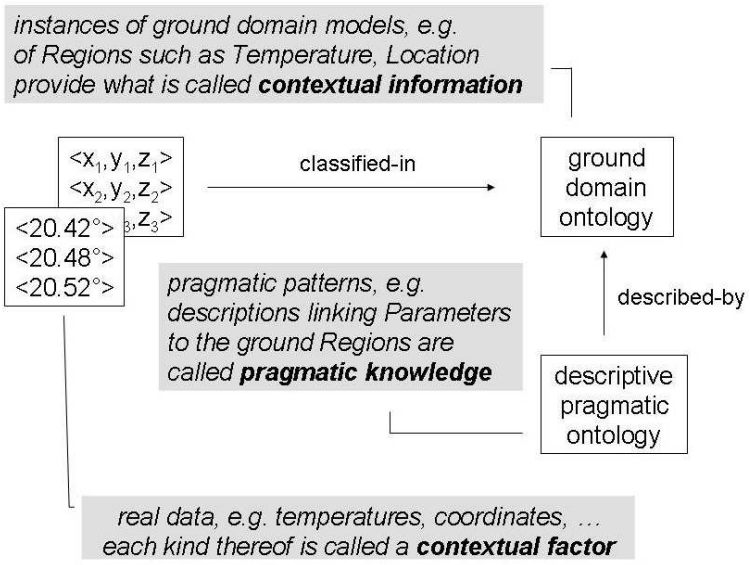

Figure 5.1: Overview of terms and referents

Motivated by this central aim, I specified ancillary aims, which I will re-examine in a final summary presented in the following section.

5.1 Aims and Contributions

As one contribution I sought to provide a hitherto missing clear distinction between contextual information and the associated pragmatic knowledge. As depicted in Figure 5.1, I have presented how observed user- and situation-specific data can be classified as contextual information in terms of the vocabulary provided by ground domain models and associated to pragmatic patterns that constitute models of pragmatic knowledge. Additionally I provided, as aimed for, a set of evaluated applications of these models within this contextual computing approach in each of the three steps of processing the user's natural language input found in dialog systems [Allen, 1987], i.e.:

- verification of hypothesis after automatic speech recognition in phonetic interpretation;
- disambiguation of word senses and relation extraction in semantic interpretation;
- decontextualization of underspecified utterances and intention recognition in contextual interpretation.

These evaluations provided evidences for the contributions that are possible by means of putting things into context as proposed herein. The individual evaluations were presented together with their respective corpus-based metrics, performance- and baseline results as well as methodological and empirical examinations regarding the issue of measuring the quality of the underlying models as a whole. Also, a descriptive ontology-based approach for enabling context-adaptive decontextualization of these interpretations was described and applied in a real time multimodal prototype system.

I intended to show herein that explicit formal knowledge models and means to observe a given context are needed to build scalable systems that seek to be able to handle context-dependent ambiguous, underspecified and noisy input. The central focus of this work, therefore, was on the development of more robust and scalable systems that can interact with human users using natural modalities, such as spoken natural language, which - as I have stated in the beginning have evolved to facilitate efficient communication among interlocutors who share vast and rich amounts of background knowledge and which is always situated in given context. In the approach taken herein what is pertinent, i.e. what contextual factors matter and which contextual information, therefore, activates a corresponding pragmatic pattern, depends on the cooperative communicative task at hand [Tomasello, 2008].

In the case of empirically learning what matters - whether by means of conducting field experiments or by evaluating the resulting prototype systems iteratively - the limitation of resources confines what it is experimentally possible within a thesis. This also means that there is additional possible and - in the light of the evidences and results presented herein - also feasible and sensible work to be undertaken, which I will discuss in the following section.

5.2 Future Work

Essentially, the work described above seeks to determine from sets of alternative forms, meanings and functions the best one by means of employing specific contextual information- and ontological knowledge sources. While I have demonstrated that - in doing so - a higher percentage of the best form, meaning and function can be identified, as compared to the individual gold standards of annotated data, there are still remaining cases where a better solution was given in the respective gold standard. In my mind resolving those cases requires an even deeper understanding of what *acting out* the resulting alternative *scenes* - that can be evoked by the situated utterance - entails and yields. For realizing this computationally, the inclusion of simulation-based approaches provide suitable instruments to enable a system to simulate each alternative and to score which of the results yielded by the set of simulations constitutes the best-fitting one, given a suitable scoring function.

I have pointed out before, such an approach has been pursued for understanding narratives and other declarative texts, e.g. newspaper articles [Narayanan, 1997, Narayanan, 1999b, Narayanan, 1999a], within the Neural The-

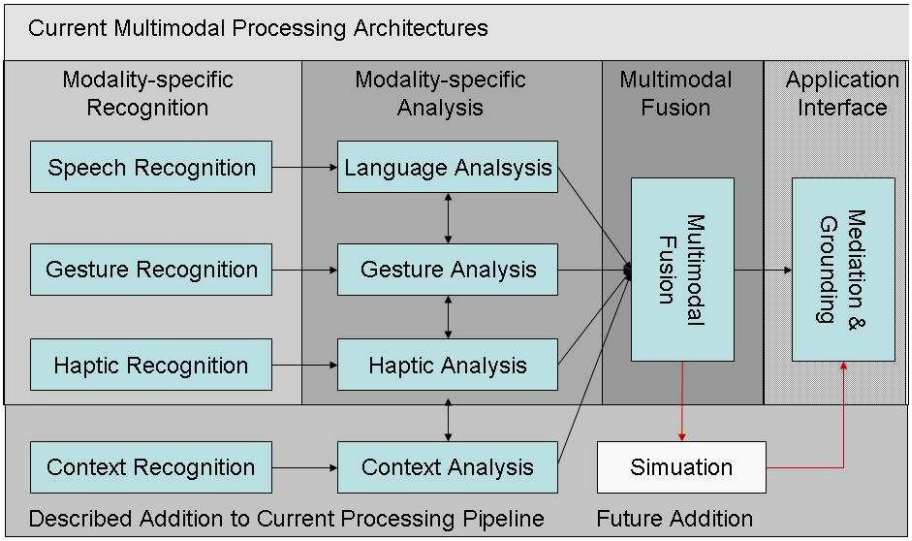

Figure 5.2: Context-aware multimodal system with simulation

ory of Language project [Feldman, 2006] and seems ideally suited to approach understanding utterances, as given in Example 42 from Corpus ASK_1 - classified therein as Type G *Other* in Table 4.2, in addition to understanding Types A - E as well.

(42) I don't see any bus stops

Given a way of determining that, in a situation S_i, it is best to construe the declarative utterance given in Example 42 as an instructional request means assuming that this most plausible construal leads to the most felicitous response. This can, as stated above, be regarded as the response that, in turn, advances the involved partners' cooperative task-specific effort closer to the desired goal state, i.e. to find the best individual response in the situated joint action at hand. This is also the approach taken in several fields of artificial intelligence, e.g. in sequential action learning [Sutton and Barto, 1998] or for incremental language and event processing [Dominey, 2007]. Akin to all of these approaches is that individual knowledge sources have to be modeled and linked to the contextual information observed. A consequential future research effort could, therefore, seek to employ the models described above as a *pragmatic specifications* for the simulations to be performed.

When using the pragmatic- and context-dependent enhancement proposed herein, the modality of context can be consequentially, be treated as the other modalities, where form instances are recognized and their meaning is analyzed with respect to the given knowledge models to yield semantic specifications thereof, as schematically depicted in Figure 5.2.

Given the contributions of treating context as a *bona fide* modality, as I have proposed and examined in this work using the corresponding representational instruments and scoring functions, one can expect that simulations can be performed based on both the resulting semantic- and pragmatic specifications. Also this feasible addition - situated in the context of multimodal dialog processing - is schematically shown in Figure 5.2.

As in all knowledge-based approaches one has to respond to the general question how the required knowledge can be obtained, for which - next to experiments of the type discussed in Section 4.2 - also additional learning experiments, beyond the standard machine learning approaches [Winston, 1992] can also be pursued in future work. The most promising paradigms for this, in my mind, are:

- the paradigm of language games [Steels, 2001] for agent- and robot-based self-organized grounded learning of linguistic forms [Steels, 2008] and cultural categories [Puglisi et al., 2008];
- the paradigm of human-based computing [Ahn, 2008] for eliciting human-made categorizations and form-meaning mappings [Ahn, 2007].[1]

In line with pursuing corresponding contextually enhanced learning and acquisition experiments, there certainly is additional work in terms of fleshing out the corresponding models and their connections. This is the case for the model of linguistic information needed for morpho-syntactic decomposition [Buitelaar et al., 2006] and its connection to a model of constructional information [Porzel et al., 2006b][2] together with their respective descriptive pragmatic patters. Let me note once more, in this respect, that combing and associating these knowledge sources in addition to the ground domain knowledge is made possible by means of using a correspondingly axiomatized foundational framework and its modules, as employed herein [Masolo et al., 2002, Oberle et al., 2007].

I will conclude this section, by stating that both the traditional as well as the novel learning approaches mentioned above are being pursued at the time of writing this work [Kahn, 2009, Takhtamysheva, 2009] yielding promising results. Lastly, there are new challenges for usability testing that arise with adaptive computing technologies, e.g. ones that enable systems to behave differently in different situations. As this exploration of how far a representational approach to contextual computing can be pushed - or, in other words, to attempt to bite representational bullet - still has to be embedded in an *embodied* interactional context, which I will sketch out in the concluding remarks given in the following section.

[1] For finding corresponding context-dependent form-function mappings one would have to extend this paradigm, for example, by adopting mobile gaming techniques for a situated human-based computing approach [Grüter and Oks, 2007, Eirund and Haalck, 2008].

[2] Employing the ontology of information objects [Guarino, 2006], which has been introduced as a DOLCE module in Section 4.3, one can model constructions logically as *InformationObjects* that *express* a *Description*.

5.3 Concluding Remarks

I will return to the larger issues concerning the so-called *representational* and *interactional* approaches to context [Dourish, 2001] presented in the beginning of Chapter 2. Therein, I summarized several properties of representational approaches as discussed in Dourish (2001). These properties will now be re-examined in terms of how they apply to the work presented herein and what this entails for situating my approach to contextual computing in the field outlined by these general distinctions in conclusive manner.

- Representational approaches see context as a form of information, i.e. context is something that can be known, represented and encoded in software systems.

In a sense, this holds true for this approach, as contextual information constitutes a key component - or informational source - in this approach. However, this approach, given the terminological distinctions depicted in Figure 5.1, clearly specifies:

- how contextual information can be known, i.e. by means of observing - being aware of - pertinent topical data,
- how observed real world data can be represented, i.e. by means of classifying it in terms of a ground domain model, and
- how it can be encoded in software systems, i.e. by means of ontology engineering approaches, e.g. using suitable design patterns.

Moreover, I have not only shown what contextual information is, but also what it is not.[3] That is, while contextual information constitutes an important source of information, additional *descriptive* knowledge is needed to interweave it with the ground domain models that are employed by the context-aware system.

- Representational approaches see context as delineable, i.e. it is thought to be possible to define what counts as context for a specific application in advance.

This is not the case in this approach, as it does not seek to define a set of rules where for each pre-defined contextual setting a specific application behavior is specified in advance. Rather than following this traditional *blueprint*, this approach provides a set of logical- and content patterns that model the observable contextual information as instances of ground domain knowledge. In addition to employing such domain models for representing domain- and discourse context as described in Chapter 3, I have also shown how it can applied as a knowledge source, by means of employing the respective cardinalities of the hitherto

[3] While Dourish (ibid) does not provide a computational definition of *information*, his examples are in the scope of what, for example, a *xmls:SimpleType* could represent.

implicit semantic relations as exemplified in Section 4.1. As described in the sections thereafter, this ground ontology was linked to a descriptive model of pragmatic knowledge, using specific content patterns, e.g. the pragmatic pattern presented in Figure 4.3 and only logical patterns provided by the foundational ontology.

Depending on the observed contextual information, these patterns may or may not become activated, i.e. put into the *ActivatedPattern*-pool, whereby the one that *fleshes out* the given representation of the user's multimodal input the most is used to explicate this representation. This can, when more than one construal is represented in the set of alternative intention hypotheses, lead to corresponding disambiguation of the underlying pragmatic ambiguity as described in Section 4.2. This fleshing out is also not pre-defined, but dependent on the aggregate set of contextual information observed[4], or - in other words - the pragmatic pattern applied is the one which is the most congruent to the context at hand.

Moreover, different types of learning approaches for obtaining the required knowledge can be applied. I have showcased a machine learning-based approach for finding what matters for finding the best-fitting construal of *Where-interrogatives* in Section 4.2, demonstrating that - in cases of pragmatic ambiguity - utterances are constructed taking the shared context into account and can, therefore, also be construed by recourse to that context. I have, additionally pointed at alternative and promising learning approaches one can pursue in Section 5.2 above.

- Representational approaches see context as stable, i.e. while context may vary from application to application, it does not vary from instance to instance of an interaction with an application.

Whether this holds for this approach or not depends on how one defines an *instance of an interaction*. If no additional context sources, domain- and pragmatic knowledge are added during run time and a question about an entity were to be posed twice - while all four types of context, presented in Table 2.7, provide the same type of information regarding, the interlocutors, their location, the state of the entity as well as regarding the prior discourse, then the identical explication would indeed be performed and the same construal would win. However likely this may be - given, for example, the discourse sensitivity that I presented both in terms of applying the ground model in Section 3.3 and terms of including domain context via the descriptive pragmatic patterns in Section 4.4 - this might not even be unwanted in the light of expected system behavior.

Nevertheless, a truly unwanted effect would be to construe an immediate repetition of an utterance as one would construe the first. Aside from devising a pattern that captures the corresponding pragmatic knowledge that a repetition can - in some contexts - mean that the a misrecognition or misconstrual

[4]Please note, that one still can pre-define required information, e.g. for all reservation processes in the ground model, using logical cardinalities, as shown in Section 4.1.4 within this approach.

took place, one could alternatively seek to classify such events differently by implementing a short term memory, e.g. a so called *reservoir* as in recurrent echo state networks [Jäeger, 2003]. Nevertheless, this property of treating context as stable holds only if an instance of an interaction is to include everything, otherwise this approach - while seeking to provide robustness and dialogical efficiency - allows contextual changes to have an effect or to not have one, i.e. a change only has an effect when enough pertinent evidences are gathered that make the assumption of a corresponding change of *frame* more plausible.

- Representational approaches see context and activity as being separable, i.e. therein context is taken to describe features of the environment within which an activity takes place but the elements of the activity do not belong to context itself.

This, certainly, does not hold true for this approach. Not only does it - as I have noted through this work - all start with a task at hand, which holds for the dialogical tasks of constructing and construing ground interrogatives which in can be described as individual interlocutionary courses of events such as *Questioning*; as discussed in Section 4.4 or as showcased in Section 4.2 - using the example of construing *Where-interrogatives* against the backdrop of a task-specific contextual frame. In other words, the given task can be said to provide the *viewport* on the contextual frame. A specific view can fade-out some factors - and their respective contextual information - or fade-in other factors - by virtue of their pertinence for the task.

For describing this one can adopt models of so-called *referential* movements in discourse [von Stutterheim and Klein, 1989] for corresponding *focal movements* in the modality of context, that is:

new - when a hitherto irrelevant factor becomes pertinent;

continuation - when a pertinent factor stays pertinent;

re-entry - when a formerly pertinent factor becomes pertinent again.

Please note that, as depicted in Figure 5.1, contextual *factor* refers to a type of topical data that can be observed and represented in types of contextual information that are computationally encapsulated as described in Section 4.4.2.

As I have sought to treat context as a *bona fide* modality with distinct types of information sources, which need to be recognized and analyzed as other modalities, one can consider them representational as discussed above. Nevertheless, I have sought to provide a computationally feasible approach, which - as in the case of the other modalities as well - requires knowledge sources that can be learned and engineered in suitable ways, as I have shown in Sections 3.1, 4.2.2 and 4.3. Analogously, as the morpho-syntactic and semantic capabilities of a system remain the same when the employed formal linguistic grammars and ground ontologies stay constant,[5] so do the pragmatic capabilities of a system when the contextual grammars and descriptive ontologies do not

[5]This holds true, of course for models acquired via offline learning approaches, e.g. ones based on data annotated by humans.

change over the course of the interaction(s). For this, however, embodied interaction principles, that combine interactional and representational instruments [Malaka and Porzel, 2009] can be employed as well as self-organized learning approaches [Baronchelli et al., 2009] that provide the consequential fluid emergence of new forms, meanings and functions.

Taken as a whole, evaluating the individual contributions of treating context as a true modality in understanding what a given situated utterance means has shown that it would be feasible to circumvent the baseline parsing systems employed herein. In doing so one would still produce at least the same amount of correctly disambiguated formal conceptualizations of the given utterance, by recourse to contextual information and the associated ground and descriptive knowledge alone. While it was not the intention of this work to build a parser that runs on context alone, but to seek to augment natural language processing, it still goes - in my mind - some ways to make the significance of context evident, which is vindicated furthermore by the significant context-specific performance results, as presented, ofr example, in Table 3.30 in Section 3.7.

As intended, however, I have taken the three levels of analyses that occur in the case natural language processing, i.e. going from *forms* via *meaning* to *function*, to supply challenging applications for this approach to contextual computing. More precisely, in each application a specific type of problem, i.e. noise, ambiguity and underspecification, was examined, as

- in the case of *forms* I have applied this approach to noisy data in the tasks of ranking and classifying speech recognition hypothesis;
- in the case of *meaning* I have applied this approach to ambiguous word senses and alternative semantic relations in the tasks of disambiguating correct concepts and extracting appropriate semantic relations;
- in the case of *function* I have applied this approach to underspecified semantic specifications in the tasks of explicating implicit information and construing pragmatically ambiguous utterances.

In all cases I have examined how this application contributes towards providing the conversational capabilities needed to enable robust, efficient and felicitous dialogical behavior for the non-human partner in what is supposed to be a cooperative joint (inter)action situated in a shared context. I hope the corresponding results povide evidences for the contribution of the contextual computing approach presented herein together with their empirical data, experimental settings as well as their theoretical and methodological foundations.

Bibliography

[Abney, 1996] Abney, S. (1996). Partial parsing via finite-state cascades. In *Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information*, pages 8–15, Prague, Czech Republic.

[Abu-Hakima et al., 1993] Abu-Hakima, S., Halasz, M., and Phan, S. (1993). An approach to hypermedia in diagnostic systems. In Maybury, M. T., editor, *Intelligent Multimedia Interfaces*, pages 225–256. AAAI Press, Menlo Park, CA.

[Ahn, 2007] Ahn, L. V. (2007). Human computation. In *K-CAP '07: Proceedings of the 4th international conference on Knowledge capture*, pages 5–6, New York, NY, USA. ACM.

[Ahn, 2008] Ahn, L. V. (2008). *Human Computing*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittburgh, PA.

[Akman and Surav, 1996] Akman, V. and Surav, M. (1996). Steps toward formalizing context. *AI Magazine*, 17(3):55–72.

[Alexandersson and Becker, 2001] Alexandersson, J. and Becker, T. (2001). Overlay as the basic operation for discourse processing. In *Proceedings of IJCAI*. Springer-Verlag.

[Alexandersson and Becker, 2003] Alexandersson, J. and Becker, T. (2003). The Formal Foundations Underlying Overlay. In *Processings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, Tilburg, The Netherlands.

[Alexandersson et al., 2006] Alexandersson, J., Becker, T., and Pfleger, N. (2006). Overlay: The basic operation for discourse processing. In Wahlster, W., editor, *SmartKom - Foundations of Multimodal Dialogue Systems*, pages 255–268. Springer Verlag.

[Alexandersson et al., 1995] Alexandersson, J., Maier, E., and Reithinger, N. (1995). A robust and efficient three-layered dialog component for a speech-to-speech translation system. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics,* Dublin, Ireland, 27–31 March 1995, pages 62–70.

[Allen, 1987] Allen, J. (1987). *Natural Language Understanding*. Ben. Cummings.

[Allen et al., 2000] Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. (2000). An architecture for a generic dialogue shell. *Natural Language Engineering*, 6(3 & 4):213–228.

[Allen et al., 2001a] Allen, J., Byron, K., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. (2001a). Towards conversational human-computer interaction. *AI Magazine*, 22(4):27–37.

[Allen et al., 2001b] Allen, J., Ferguson, G., and Stent, A. (2001b). An architecture for more realistic conversational systems. In *Proceedings of Intelligent User Interfaces*, pages 14–17, Santa Fe, NM.

[Allen et al., 1996] Allen, J. F., Miller, B., Ringger, E., and Sikorski, T. (1996). A robust system for natural spoken dialogue. In *Proc. of ACL-96*.

[Allen and Perrault, 1986] Allen, J. F. and Perrault, C. R. (1986). Analyzing intention in utterances. In Grosz, B. J., Sparck Jones, K., and Webber, B. L., editors, *Natural Language Processing*, pages 441–458. Kaufmann, Los Altos, CA.

[Allen et al., 1995] Allen, J. F., Schubert, L. K., Ferguson, G., Heeman, P., Hwang, C. H., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M., and Traum, D. (1995). The TRAINS project: A case study in building a conversational agent. *Journal of Experimental and Theoretical AI*, 7:7–48.

[Alshawi and Moore, 1992] Alshawi, H. and Moore, R. C. (1992). Introduction to the cle. In Alshawi, H., editor, *The Core Language Engine*, pages 1–10. MIT Press, Cambridge, MA.

[Ammann, 1928] Ammann, H. (1928). *Die menschliche Rede. Sprachphilosophische Untersuchungen. 2. Der Satz*. Lahr, Schauenburg.

[Anderson et al., 1993] Anderson, H. T. A., Bard, E., Doherty-Sneddon, G., Newlands, A., and Sotillo, C. (1993). The hcrc map task corpus: natural dialogue for speech recognition. In *HLT '93: Proceedings of the workshop on Human Language Technology*, pages 25–30, Morristown, NJ, USA. Association for Computational Linguistics.

[Anderson et al., 1995] Anderson, J. R., Corbett, A. T., Koedinger, K. R., and Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2):167–207.

[André, 1999] André, E. (1999). Towards personalized multimedia presentation systems for all. In *HCI (2)*, pages 873–877.

[Angeles, 1981] Angeles, J. (1981). *Dictionary of Philosophy*. Harper Perennial, New York, NY.

[Ankolekar et al., 2006] Ankolekar, A., Buitelaar, P., Cimiano, P., Hitzler, P., Kiesel, M., Krtzsch, M., Lewen, H., Neumann, G., Sintek, M., Tserendorj, T., and Studer, R. (2006). Smartweb: Mobile access to the semantic web. In *Proc. of the Demo Session at the International Semantic Web Conference*, Athens GA, USA.

[Aust et al., 1995] Aust, H., Oerder, M., Seide, F., and Steinbiss, V. (1995). The Philips automatic train timetable information system. *Speech Communication*, 17:249–262.

[Austin, 1962] Austin, J. L. (1962). *How To Do Things With Words*. Oxford University Press, Oxford U.K.

[Baader et al., 2003] Baader, F., Horrocks, I., and Sattler, U. (2003). Description logics as ontology languages for the semantic web. In *Festschrift in honor of Jrg Siekmann, Lecture Notes in Artificial Intelligence*, pages 228–248. Springer-Verlag.

[Baader et al., 2006] Baader, F., Lutz, C., and Suntisrivaraporn, B. (2006). Cel: A polynomial-time reasoner for life science ontologies. In Furbach, U. and Shankar, N., editors, *Proceedings of the 3rd International Joint Conference on Automated Reasoning (IJCAR'06)*, volume 4130 of *Lecture Notes in Artificial Intelligence*, pages 287–291. Springer-Verlag.

[Babitski, 2004] Babitski, G. (2004). *Inferenzalgorithmen zur Auswahl ontologiebasierter Situationsbeschreibungen fr ein kontextadaptives Dialogsystem*. Diploma Thesis at the Computer Science Departement of the Technical University of Darmstadt.

[Baker et al., 1998] Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet P roject. In *Proceedings of COLING-ACL*, Montreal, Canada.

[Baronchelli et al., 2009] Baronchelli, A., Cattuto, C., Loreto, V., and Puglisi, A. (2009). Complex systems approach to the emergence of language. In Minett, J. and Wang, W., editors, *Language, Evolution and the Brain*, page in press. City University of Hong Kong Press, Hong Kong, China.

[Barr-Hillel, 1954] Barr-Hillel, Y. (1954). Logical Syntax and Semantics. *Language*, 20.

[Barwise and Perry, 1983] Barwise, J. and Perry, J. R. (1983). *Situations and Attitudes*. MIT Press, Cambridge, Mass.

[Bateman and Zock, 2003] Bateman, J. and Zock, M. (2003). Natural language generation. In Mitkov, R., editor, *Oxford Handbook of Computational Linguistics*, chapter 15, pages 284–304. Oxford University Press, Oxford.

[Bazire and Brézillon, 2005] Bazire, M. and Brézillon, P. (2005). Understanding context before using it. In Dey, A. K., Kokinov, B. N., Leake, D. B., and Turner, R. M., editors, *5th International and Interdisciplinary Conference on Modeling and Using Context*, volume 3554 of *Lecture Notes in Computer Science*, pages 29–40. Springer.

[Bergen, 2001] Bergen, B. (2001). *Of sound, mind, and body: neural explanations for non-categorical phonology.* PhD thesis, UC Berkeley.

[Beringer, 2003] Beringer, N. (2003). The SmartKom Multimodal Corpus - Data Collection and End-to-End Evaluation. In *Colloquium of the Department of Linguistics*, University of Nijmwegen.

[Beringer et al., 2002] Beringer, N., Kartal, U., Louka, K., Schiel, F., and Türk, U. (2002). PROMISE: A Procedure for Multimodal Interactive System Evaluation. In *Proceedings of the Workshop 'Multimodal Resources and Multimodal Systems Evaluation*, Las Palmas, Spain.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, May.

[Berton et al., 2006] Berton, A., Kaltenmeier, A., Haiber, U., and Schreiner, O. (2006). Speech recognition. In Wahlster, W., editor, *SmartKom - Foundations of Multimodal Dialogue Systems*, pages 85–108. Springer Verlag.

[Blohm et al., 2007] Blohm, S., Cimiano, P., and Stemle, E. (2007). Harvesting relations from the web -quantifiying the impact of filtering functions. In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI-07)*, pages 1316–1323. Association for the Advancement of Artificial Intelligence (AAAI).

[BNC, 2008] BNC (2008). The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/ (last accessed: 10/10/2008).

[Boves, 2004] Boves, L. (2004). Robust conversational system design. In *Proceedings, COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, Norwich, 30-31 August.

[Boy, 1991] Boy, G. A. (1991). On-line user model acquisition in hypertext documentation. In *Proc. of the IJCAI-91 Workshop 'Agent Modelling for Intelligent Interaction'*, pages 34–42, Sydney, Australia.

[Branigan and Pearson, 2006] Branigan, H. and Pearson, J. (2006). Alignment in Human-Computer Interaction. In Fischer, K., editor, *How People Talk to Computers, Robots, and Other Artificial Communication Partners: Proceedings of the Workshop Hansewissenschaftskolleg*, pages 140 –156.

[Brennan, 1996] Brennan, S. (1996). Lexical entrainment in spontaneous dialogue. In *Proceedings of the International Symposium on Spoken Dialogue*, pages 41–44, Philadelphia, USA.

[Brennan, 1998] Brennan, S. (1998). Centering as a psychological resource for achieving joint reference in spontaneous discourse. In Walker, M., Joshi, A., and Prince, E., editors, *Centering in Discourse*, pages 227–249. Oxford University Press, Oxford, U.K.

[Brennan, 2000] Brennan, S. (2000). Processes that shape conversation and their implications for computational linguistics. In *Proceedings of ACL*, Hong Kong.

[Brewster et al., 2004] Brewster, C., Alani, H., Dasmahapatra, S., and Milks, Y. (2004). Data driven ontology evaluation. In *Proceedings of LREC 2004*, Lisbon, Portugal. To appear.

[Bronnenberger et al., 1997] Bronnenberger, W., Bunt, H., Landsbergen, J., Medema, P., Scha, R., and Schoenmakers, W. (1997). The question-answering system phliqa1. In Bolc, L., editor, *In Natural Communication with Computers*. Carl Hanser, Munich and Macmillan, London.

[Bryant, 2003] Bryant, J. (2003). Smantic analysis with ecg. In *Proceedings of 8th International Conference on Cognitive Linguistics*.

[Bryant, 2008] Bryant, J. (2008). *Best-fit Constructional Analysis*. PhD thesis, Computer Science Division, EECS Department, University of California at Berkeley.

[Bryant et al., 2001] Bryant, J., Chang, N., Porzel, R., and Sanders, K. (2001). Where is natural language understanding. In *Proceedings of the 7th International Conference 7th Annual Conference on Architectures and Mechanisms for Language Processing*.

[Bub and Schwinn, 1999] Bub, T. and Schwinn, J. (1999). The verbmobil prototype system – a software engineering perspective. *Natural Language Engineering*, 5(1):95–112.

[Buitelaar et al., 2006] Buitelaar, P., Declerck, T., Frank, A., Racioppa, S., Kiesel, M., Sintek, M., Engel, R., Romanelli, M., Sonntag, D., Loos, B., Micelli, V., Porzel, R., and Cimiano, P. (2006). Linginfo: Design and applications of a model for the integration of linguistic information in ontologies. In *In Proceedings of the OntoLex Workshop at LREC, pp. 28 - 32. May, Genoa, Italy,.*

[Buitelaar and Magnini, 2005] Buitelaar, P. and Magnini, P. C. B., editors (2005). *Ontology learning from text : methods, evaluation and applications*, volume 123 of *Frontiers in artificial intelligence and applications*. IOS Press.

[Bunt, 1984] Bunt, H. (1984). The resolution of quantificational ambiguity in the tendum system. In *Proc. of the 10th COLING*, pages 130–133, Stanford, CA.

[Bunt, 2000] Bunt, H. (2000). Dialogue Pragmatics and Context Specification. In *Computational Pragmatics, Abduction, Belief and Context*. John Benjamins.

[Bunt, 1989] Bunt, H. C. (1989). Information dialogues as communicative action in relation to partner modelling and information processing. In Taylor, M. M., Neel, F., and Bouwhuis, D. G., editors, *The Structure of Multimodal Dialogue*, pages 47–73. North-Holland, Amsterdam.

[Bürkle, 1986] Bürkle, B. (1986). "von mir aus ..." Zur hörerbezogenen lokalen Referenz. Technical Report Bericht 10, Forschergruppe Sprachen und Sprachverstehen im sozialen Kontext.

[Byron, 2002] Byron, D. K. (2002). Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 80–87.

[Cahour and Karsenty, 1993] Cahour, B. and Karsenty, L. (1993). Context of dialogue: a cognitive point of view. Technical Report 93/13, LAFORIA, University Paris 6.

[Carletta, 1996] Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

[Carpenter, 1992] Carpenter, B. (1992). *The Logic of Typed Feature Structures*. Cambridge University Press, Cambridge.

[Cassell, 2001] Cassell, J. (2001). Embodied Conversational Agents: Representation and Intelligence in User Interface. *AI Magazine*, 22(4):67–84.

[Cettolo et al., 1999] Cettolo, M., Corazza, A., Lazzari, G., Pianesi, F., Pianta, E., and Tovena, L. M. (1999). A speech-to-speech translation based interface for tourism. In *In Proceedings of the ENTER Conference*.

[Chang et al., 2002] Chang, N., Feldman, J., Porzel, R., and Sanders, K. (2002). Scaling cognitive linguistics: Formalisms for language understanding. In *Proceedings of the 1st International Workshop on Scalable Natural Language Understanding (ScaNaLU)*, Heidelberg, Germany.

[Chin, 1993] Chin, D. (1993). Aquiring user-models. *Artificial Intelligence Review*, 7:186–197.

[Chomsky, 1965] Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Mass.

[Chomsky, 1981] Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

[Chomsky, 1995] Chomsky, N. (1995). *The Minimalist Program*. MIT Press, Cambridge, Mass.

[Chotimongcol and Rudnicky, 2001] Chotimongcol, A. and Rudnicky, A. (2001). N-best speech hypotheses reordering using linear regression. In *Proceedings of Eurospeech*, pages 1829–1832, Aalborg, Denmark.

[Chuang, 300] Chuang, T. (300). True classic of southern (cultural) florescence.

[Ciaramita et al., 2005] Ciaramita, M., Gangemi, A., Ratsch, E., Rojas, I., and Saric, J. (2005). Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI2005)*. Edinburgh, UK.

[Cimiano et al., 2004] Cimiano, P., Eberhart, A., Hitzler, Pascal Oberle, D., Staab, S., and Studer, R. (2004). The smartweb foundational ontology. *SmartWeb Project Report*.

[Cimiano et al., 2005] Cimiano, P., Pivk, A., Schmidt-Thieme, L., and Staab, S. (2005). Learning taxonomic relations from heterogeneous sources of evidence. In Buitelaar, P., Cimiano, P., and Magnini, B., editors, *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123 of *Frontiers in Artificial Intelligence*, pages 59–73. IOS Press.

[Clark and Carlson, 1981] Clark, H. and Carlson, T. (1981). Context for comprehension. In Long, J. and Beddeley, A., editors, *Attention and Performace*. Cambridge University Press.

[Clark and Marshall, 1981] Clark, H. H. and Marshall, C. (1981). Definite reference and mutual knowledge. In Joshi, A., Webber, B., and Sag, I., editors, *Linguistic Structure and Discourse Setting*. Cambridge University Press.

[Cobuild, 1995] Cobuild (1995). *Collins COBUILD English Dictionary*. Harper Collins, London.

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

[Cohen et al., 1997] Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., and Clow, J. (1997). Quickset: Multimodal interaction for distributed applications. In *Proceedings of the Fifth Annual International ACM Multimedia Conference,* Seattle, Association for Computing Machinery. http://www.cse.ogi.edu/CHCC/Papers/philPaper/acm.html.

[Connolly, 2001] Connolly, J. H. (2001). Context in the study of human languages and computer programming languages: A comparison. *Modeling and Using Context*, Springer, LNCS:116–128.

[Coors et al., 2000] Coors, V., Kray, C., and Porzel, R. (2000). Zu komplexen diensten mit einfachen natürlichsprachlichen interaktionen. In *Workshop on Digital Storytelling (DISTEL 2000)*, Darmstadt, Germany.

[Cormen et al., 1990] Cormen, T. H., Leiserson, C. E., and Rivest, R. R. (1990). *Introduction to Algorithms*. MIT press, Cambridge, MA.

[Cozman, 1998] Cozman, F. (1998). Interchange format for bayesian networks. http://www.cs.cmu.edu/ fgcozman/Research/InterchangeFormat/ (last accessed: 04/16/2009.

[Cozman, 2000] Cozman, F. (2000). Generalizing variable elimination in Bayesian networks. In *Proceedings of the IBERAMIA Workshop on Probabilistic Reasoning in Artificial Intelligence*, Sao Paulo, Brazil.

[Crysmann et al., 2002] Crysmann, B., Frank, A., Bernd, K., Mueller, S., Neumann, G., Piskorski, J., Schaefer, U., Siegel, M., Uszkoreit, H., Xu, F., Becker, M., and Krieger, H.-U. (2002). An integrated archictecture for shallow and deep processing. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[Darves and Oviatt, 2002] Darves, C. and Oviatt, S. (2002). Adaptation of Users' Spoken Dialogue Patterns in a Conversational Interface. In *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, U.S.A.

[Deemter et al., 1985] Deemter, K., Brockhof, G., Bunt, H., Meya, M., and de Vet, J. (1985). From tendum to spicos, or: How flexible is the tendum approach to question answering? Technical Report APR 20, IPO.

[Demetriou and Atwell, 1994] Demetriou, G. and Atwell, E. (1994). A semantic network for large vocabulary speech recognition. In Evett, L. and Rose, T., editors, *Proceedings of AISB workshop on Computational Linguistics for Speech and Handwriting Recognition*, University of Leeds.

[Dey, 2001] Dey, A. K. (2001). Understanding and using context. *Personal and Ubiquitous Computing*, 5:4–7.

[Dey and Mankoff, 2005] Dey, A. K. and Mankoff, J. (2005). Designing mediation for context-aware applications. *ACM Trans. Comput.-Hum. Interact.*, 12(1):53–80.

[Dijkstra, 1959] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271.

[Ding, 2005] Ding, Z. (2005). *BayesOWL: A Probabilistic Framework for Semantic Web*. PhD thesis, University of Maryland.

[Dix et al., 2004] Dix, A. J., Finlay, J., and Abowd, G. D. (2004). *Human-computer interaction*. Pearson Prentice-Hall, Harlow, 3. ed. edition.

[Dominey, 2007] Dominey, P. F. (2007). Towards a construction-based framework for development of language, event perception and social cognition: Insights from grounded robotics and simulation. *Neurocomputing*, 70(13-15):2288–2302.

[Dourish, 2001] Dourish, P. (2001). *Where the Action Is: The Foundations of Embodied Interaction*. MIT Press.

[Dourish, 2004] Dourish, P. (2004). What We Talk About When We Talk About Context. *Personal and Ubiquitous Computing*, 8(1):19–30.

[Dourish, 2007] Dourish, P. (2007). Responsibilities and implications: Further thoughts on ethnography and design. In *Proc. ACM Conf. Designing for the User Experience DUX 2007*, Chicago, IL.

[Duncan, 1974] Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 3.

[Ebert et al., 2001] Ebert, C., Lappin, S., Gregory, H., and Nicolov, N. (2001). Generating full paraphrases out of fragments in a dialogue interpretation system. In *Proceedings 2nd SIGdial Workshop*, pages 58–67, Aalborg, Denmark.

[Eco, 1984] Eco, U. (1984). *The Role of the Reader: Explorations in the Semiotics of Texts*. Indiana University Press, Bloomington, IN.

[Edmonds, 2002] Edmonds, P. (2002). SENSEVAL: The evaluation of word sense disambiguation systems. *ELRA Newsletter*, 7/3.

[Eirund and Haalck, 2008] Eirund, H. and Haalck, M. (2008). Context aware gaming auf mobiltelefonen im spiel fanmob. In Herczeg, M. and Kindsmüller, M. C., editors, *Mensch und Computer*, pages 397–400. Oldenburg Verlag.

[Elting, 2002] Elting, C. (2002). What are Multimodalities made of? Modeling Output in a Multimodal Dialog System. In *ISAMP'02 Workshop on Intelligent Situation-Aware Media and Presentations*, Edmonton, Canada.

[Elting et al., 2002] Elting, C., Zwickel, J., and Malaka, R. (2002). Device-dependant modality selection for user-interfaces - an empirical study. In *International Conference on Intelligent User Interfaces*, San Francisco, CA.

[EMMA, 2004] EMMA (2004). EMMA: Extensible multimodal annotation markup language. http://www.w3.org/TR/2004/WD-emma-20041214/ (last accessed: 04/18/2009).

[Engel, 2002] Engel, R. (2002). SPIN: Language understanding for spoken dialogue systems using a production system approach. In *Proceedings of ICSLP 2002*.

[Fauconnier, 1985] Fauconnier, G. (1985). *Mental Spaces: Aspects of Meaning Construction in Natural Language*. MIT Press/Bradford, Cambridge, Mass. and London.

[Faucounnier and Turner, 1998] Faucounnier, G. and Turner, M. (1998). Conceptual integration networks. *Cognitive Science*, 22:133–187.

[Feldman et al., 1996] Feldman, J., Lakoff, G., Bailey, D., Narayanan, S., Regier, T., and Stolcke, A. (1996). l_0—the first five years of an automated language acquisition project. *AI Review*, 10:103–129.

[Feldman, 2006] Feldman, J. A. (2006). *From Molecule to Metaphor: A Neural Theory of Language*. MIT Press, Cambridge, MA.

[Fensel et al., 2001] Fensel, D., van Harmelen, F., Horrocks, I., McGuinness, D., and Patel-Schneider, P. (2001). Oil: An ontology infrastructure for the semantic web. *IEEE Intelligent Systems*, 16(2).

[Ferguson and Allen, 1998] Ferguson, G. and Allen, J. F. (1998). TRIPS: An intelligent integrated problem-solving assistant. In *Proceedings of the 15th National Conference on Artificial Intelligence & 10th Conference on Innovative Applications of Artificial Intelligence,* Madison, Wisc., 26–30 July 1998, pages 567–573.

[Ferguson et al., 1996] Ferguson, G., Allen, J. F., Miller, B., and Ringger, E. (1996). The desgin and implementation of the trains-96 system. Technical Report 96-5, University of Rochester, New York.

[Fetter, 1998] Fetter, P. (1998). Detection and transcription of OOV words.

[Fillmore, 1988] Fillmore, C. (1988). The mechanisms of construction grammar. In *Berkeley Linguistics Society*, volume 14, pages 35–55.

[Fillmore and Baker, 2000] Fillmore, C. J. and Baker, C. F. (2000). FrameNet: Frame semantics meets the corpus. In *Linguistic Society of America*.

[Fischer, 2006] Fischer, K. (2006). How people really talk to computers. In Fischer, K., editor, *Proccedings of the Workshop on How People Talk to Computers, Robots and other Artificial Agents*, pages 47–73. SFB, Bremen.

[Fodor, 1983] Fodor, J. A. (1983). *The Modularity of Mind*. MIT Press, Cambridge, MA.

[Foltz et al., 1998] Foltz, P., Kintsch, W., and Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2&3):285–307.

[Francony et al., 1992] Francony, J.-M., Kuijpers, E., and Polity, Y. (1992). Towards a methodology for wizard of oz experiments. In *Third Conference on Applied Natural Language Processing*, Trento, Italy.

[Furnas et al., 1987] Furnas, G., Landauer, T., and Dumais:, G. (1987). The vocabulary problem in human-system-communication: an analysis and a solution. *Communications of the ACM*, 30(11):964–971.

[Gallwitz et al., 1998] Gallwitz, F., Aretoulaki, M., Boros, M., Haas, J., Harbeck, S., Huber, R., Niemann, H., and Nöth, E. (1998). The Erlangen spoken dialogue system EVAR: A state-of-the-art information retrieval system. In *Proceedings of 1998 International Symposium on Spoken Dialogue (ISSD 98),* Sydney, Australia, 30. Nov., 1998, pages 19–26.

[Galton, 1892] Galton, F. (1892). *Finger Prints*. F.R.S.

[Gangemi, 2005] Gangemi, A. (2005). Ontology design patterns for semantic web content. In *M. Musen et al. (eds.): Proceedings of the Fourth International Semantic Web Conference*. Berlin, Springer.

[Gangemi et al., 2004] Gangemi, A., Borgo, S., Catenacci, C., and Lehmann, J. (2004). Metokis deliverable d07: Task taxonomies for knowledge content. http://metokis.salzburgresearch.at/files/deliverables/ (last accessed: 04/18/2009).

[Gangemi et al., 2005] Gangemi, A., Catenacci, C., Ciaramita, M., and Lehmann, J. (2005). A theoretical framework for ontology evaluation.

[Gangemi et al., 2001] Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. (2001). Understanding top-level ontological distinctions.

[Gangemi et al., 2002] Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. (2002). Sweetening ontologies with dolce.

[Gangemi and Mika, 2003] Gangemi, A. and Mika, P. (2003). Understanding the semantic web through descriptions and situations. In *Proceedings of the ODBASE Conference*. Springer.

[Garrod and Anderson, 1987] Garrod, S. and Anderson, A. (1987). Saying what you mean in dialog: A study in conceptual and semantic co-ordination. *Cognition*, 27.

[Gavalda, 1999] Gavalda, M. (1999). SOUP: A parser for real-world spontaneous speechgrowing semantic grammars. In *Proceedings of the 6th International Workshop on Parsing Technologies*, Trento, Italy.

[Gedigian et al., 2006] Gedigian, M., Bryant, J., Narayanan, S., and Ciric, B. (2006). Catching metaphors. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, pages 41–48, New York City, New York. Association for Computational Linguistics.

[Gil and Ratnakar, 2002] Gil, Y. and Ratnakar, V. (2002). A comparison of (semantic) markup languages. In *Proc.of the 15th Int.FLAIRS Conference*, Florida.

[Gildea and Jurafsky, 2002] Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

[Givón, 1995] Givón, T. (1995). *Functionalism and Grammar*. John Benjamins, Amsterdam, The Netherlands.

[Glatz et al., 1995] Glatz, D., Klabunde, R., and Porzel, R. (1995). Towards the generation of preverbal messages for spatial descriptions. Technical Report 91, SFB 245, University of Heidelberg and Mannheim, Germany.

[Goldberg, 1995] Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

[Gomez-Perez, 1999] Gomez-Perez, A. (1999). Evaluation of taxonomic knowledge in ontologies and knowledge bases. In *Proceedings of the 12th Bannf Knowledge Acquisition for Knowledge-based Systems Workshop, Banff, Canada*.

[Gorin et al., 1997] Gorin, A. L., Riccardi, G., and Wright, J. H. (1997). How may I help you? *Speech Communication*, 23:113–127.

[Grosz et al., 1977] Grosz, B. J., Hendrix, G. G., and Robinson, A. E. (1977). Using process knowledge in understanding task-oriented dialogs. In *Proc. of the 5th IJCAI*, page 90, Cambridge, MA.

[Gruber, 1993] Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition (5)*.

[Grüter and Oks, 2007] Grüter, B. and Oks, M. (2007). Situated play and mobile gaming. In Akira, B., editor, *Situated Play*, pages 103–112, Tokyo. The University of Tokyo.

[Guarino, 1998] Guarino, N. (1998). Some ontological principles for designing upper level lexical resources. In *Proceedings of the 1st International Conference on Language Resources and Evaluation,*.

[Guarino, 2006] Guarino, N. (2006). Wonderweb deliverable d2: The ontology of information objects. http://www.loa-cnr.it/Papers/Deliverable%202.pdf (last accessed: 04/18/2009).

[Guarino and Poli, 1995] Guarino, N. and Poli, R. (1995). Formal ontology in conceptual analysis and knowledge representation. *Special issue of the International Journal of Human and Computer Studies*, 43.

[Guarino and Welty, 2000] Guarino, N. and Welty, C. (2000). A formal ontology of properties. In R., D. and O., C., editors, *Proceedings of EKAW-2000: The 12th International Conference on Knowledge Engineering and Knowledge Management*, volume 1937, pages 97–112. Springer Verlag.

[Guarino and Welty, 2002] Guarino, N. and Welty, C. A. (2002). Evaluating ontological decisions with ontoclean. *Communications of the ACM*, 45(2):61–65.

[Gurevych and Porzel, 2003] Gurevych, I. and Porzel, R. (2003). Using knowledge-based scores for identifying best speech recognition hypothesis. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, pages 77–81, Chateau-d'Oex-Vaud, Switzerland.

[Gurevych et al., 2006] Gurevych, I., Porzel, R., and Malaka, R. (2006). Modeling domain knowledge: Know-how and know-what. In Wahlster, W., editor, *SmartKom - Foundations of Multimodal Dialogue Systems*, pages 71–84. Springer Verlag.

[Gurevych et al., 2003a] Gurevych, I., Porzel, R., and Merten, S. (2003a). Generating Interfaces from Ontologies. In *Proceedings of the HLT/NAACL SEALTS Workshop*, page (in press), Edmonton, Canada.

[Gurevych et al., 2003b] Gurevych, I., Porzel, R., Slinko, E., Pfleger, N., Alex, J., and Merten, S. (2003b). Less is More: Using a single knowledge representation in dialogue systems. In *In Proceedings of the HLT-NAACL03 Workshop on Text Meaning*, pages 14–21.

[Gurevych et al., 2002] Gurevych, I., Strube, M., and Porzel, R. (2002). Automatic classification of speech recognition hypothesis. In *Proceedings 3rd SIGdial Workshop,* Philadelphia, USA, July 2002, pages 90–95.

[Haarslev and Möller, 2003] Haarslev, V. and Möller, R. (2003). Racer: A core inference engine for the semantic web. In *Proceedings of the 2nd International Workshop on Evaluation of Ontology-based Tools (EON2003), located at the 2nd International Semantic Web Conference ISWC 2003, Sanibel Island, Florida, USA, October 20*, pages 27–36.

[Hahn and Amtrup, 1996] Hahn, W. V. and Amtrup, J. W. (1996). Speech-to-speech translation: The project verbmobil. In *In Proceedings of SPECOM 96*, pages 51–56.

[Hayes, 1999] Hayes, B. (1999). The web of words. *American Scientist*, 87(2).

[Hearst, 1992] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *In Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545.

[Hendrix, 1977] Hendrix, G. (1977). The lifer manual: A guide to building pratical natural language interfaces. Technical Report T-138, SRI International, Menlo Park, CA.

[Herfet et al., 2001] Herfet, T., Kirste, T., and Schnaider, M. (2001). Electronic multimodal operation and service-assistance - lead project embassi. *i-com*, 1(1):12–.

[Herrmann and Grabowski, 1994] Herrmann, T. and Grabowski, J. (1994). *Sprechen. Psychologie der Sprachproduktion.* Spektrum Akademischer Verlag.

[Higashinaka et al., 2002] Higashinaka, R., Miyazaki, N., Nakano, M., and Aikawa, K. (2002). A method for evaluating incremental utterance understanding in spoken dialogue systems. In *Proceedings of the International Conference on Speech and Language Processing 2002*, pages 829–833, Denver, USA.

[Higashinaka et al., 2003] Higashinaka, R., Miyazaki, N., Nakano, M., and Aikawa, K. (2003). Evaluating discourse understanding in spoken dialogue systems. In *Proceedings of Eurospeech*, pages 1941–1944, Geneva, Switzerland.

[Hirschman and Thompson, 1997] Hirschman, L. and Thompson, H. (1997). Overview of evaluation in speech and natural language. In Cole, R., editor, *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Cambridge.

[Hobbs, 1991] Hobbs, J. (1991). Metonomy and syntax. Technical Report 500, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025.

[Hobbs et al., 1990] Hobbs, J., Stickel, M., Appelt, D., and Martin, P. (1990). Interpretation as abduction. Technical Report 499, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025.

[Hobbs et al., 1988] Hobbs, J. R., Stickel, M. E., Martin, P., and Edwards, D. (1988). Interpretation as abduction. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics,* Buffalo, N.Y., 7–10 June 1988, pages 95–103.

[Horrocks, 1998] Horrocks, I. (1998). The FaCT system. In de Swart, H., editor, *Proc. of the 2nd Int. Conf. on Analytic Tableaux and Related Methods (TABLEAUX'98)*, volume 1397 of *Lecture Notes in Artificial Intelligence*, pages 307–312. Springer.

[Hovy, 2001] Hovy, E. (2001). Comparing sets of semantic relations. In Green, R., Bean, C., and Myaeng, S. H., editors, *Semantics of Relations*. Kluwer Academic Publishers, Dordrecht, NL.

[Ide and Veronis, 1998] Ide, N. and Veronis, J. (1998). Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, 24(1):1–40.

[Jäeger, 2003] Jäeger, H. (2003). Adaptive nonlinear system identification with echo state networks. In Becker, . T. and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 593–600. MIT Press, Cambridge, MA.

[Jameson and Wahlster, 1982] Jameson, A. and Wahlster, W. (1982). User modelling in anaphora generation: Ellipsis and definite description. In *Proceedings of the European Conference on Artificial Intelligence (ECAI '82), 1982*, pages 222–227.

[Jöest et al., 2005] Jöest, M., Häußler, J., Merdes, M., and Malaka, R. (2005). Multimodal interaction for pedestrians: an evaluation study. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pages 59–66, New York, USA. ACM.

[Johnson, 1998] Johnson, P. (1998). Usability and Mobility; Interactions on the Move. In Proceedings of the First Workshop on Human Computer Interaction with Mobile Devices. Technical Report G98-1., Department of Computing Science, University of Glasgow, Scotland.

[Johnston et al., 2002] Johnston, M., Bangalore, S., Vasireddy, G., Stent, A., Ehlen, P., Walker, M., Whittaker, S., and Maloor, P. (2002). MATCH: An architecture for multimodal dialogue systems. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 376–383.

[Jurafsky and Martin, 1991] Jurafsky, D. and Martin (1991). *NAtural Language Processing*. Springer.

[Kahn, 2009] Kahn, A. (2009). *Collecting Context Information for Mobile Computing*. Master Thesis at the University of Applied Sciences Bremen.

[Kaiser et al., 2006] Kaiser, M., Mögele, H., and Schiel, F. (2006). Bikers accessing the web: The smartweb motorbike corpus. In *Proceedings of the LREC Conference*, Genova, Italy.

[Katz, 1980] Katz, J. (1980). *A Study of Propositional Structure and Illocutionary Force*. Harvard University Press, Cambridge, Mass.

[Kay and Fillmore, 1999] Kay, P. and Fillmore, C. (1999). Grammatical constructions and linguistic generalizations: the *What's X doing Y?* construction. *Language*, 75(1):1–33.

[Kehler, 2002] Kehler, A. (2002). *Coherence, Reference and the theory of Grammar*. CSLI.

[Kilgarriff, 1993] Kilgarriff, A. (1993). Dictionary Word Sense Distinctions. *Computers and the Humanities*, 26:365–87.

[Kingsbury, 1968] Kingsbury, D. (1968). *Unpublished Honor Thesis*. PhD thesis, Havard University.

[Klabunde et al., 1999] Klabunde, R., , Glatz, D., and Porzel, R. (1999). An anatomy of a spatial description. In Meyer-Klabunde, R. and von Stutterheim, C., editors, *Representation and Processes in Natural Language Production*, pages 89–116. Deutscher Universitäts Verlag.

[Klabunde and Porzel, 1998] Klabunde, R. and Porzel, R. (1998). Tailoring spatial descriptions to the addressee: A constraint based approach. *Linguistics*, 36(3):551–577.

[Klein et al., 2000] Klein, M., Fensel, D., van Harmelen, F., and Horrocks, I. (2000). The relation between ontologies and schema-languages: translating oil-specifications in xml-schema. In *Proceedings of the Workshop on Application of Ontologies and Problem solving Methods*, Berlin, Germany.

[Kopp et al., 2004] Kopp, S., Sowa, T., and Wachsmuth, I. (2004). Imitation Games with an Artificial Agent: From Mimicking to Understanding Shape-Related Iconic Gestures. In Camurri, A. and Volpe, G., editors, *Gesture-Based Communication in Human-Computer Interaction, International Gesture Workshop, Genua, Italy April 2003*, LNAI 2915, pages 436–447. Springer.

[Kraemer et al., 2007] Kraemer, N., Simons, N., and Kopp, S. (2007). The effects of an embodied conversational agent's nonverbal behavior on user's evaluation and behavioral mimicry. In *In Proc. of 7th Conference on Intelligent Virtual Agents (IVA 07)*, pages 238–251, LNAI 4722, Springer-Verlag.

[Krauss, 1987] Krauss, R. (1987). The role of the listener. *Journal of Language and Social Psychology*, 6:81–98.

[Krauss and Weinheimer, 1964] Krauss, R. and Weinheimer, S. (1964). Changes in the lenght of reference phrases as a function of social interaction: A preliminary study. *Psychonomic Science*, (1):113–114.

[Krauss et al., 1977] Krauss, R., Weinheimer, S., and More, S. (1977). The role of audible and visible back-channel responses in interpersonal communication. *Journal of Personality and Social Psychology*, (9):523–529.

[Kray and Porzel, 2000] Kray, C. and Porzel, R. (2000). Spatial cognition and natural language interfaces in mobile personal assistans. In *Proceedings of the ECAI Workshop on Artifical Intelligence in Mobile Systems*, Berlin, Germany.

[Lakoff, 1987] Lakoff, G. (1987). *Women, Fire, and Dangerous Things. What Categories Reveal about the Mind*. Chicago University Press, Chicago, Ill.

[Lakoff and Johnson, 1980] Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.

[Lambrecht, 1994] Lambrecht, K. (1994). *Information Structure and Sentence Form: Topic, Focus, and the Mental Representation of Discourse Referents*. Cambridge University Press, Cambridge, U.K.

[Langacker, 1987] Langacker, R. W. (1987). *Foundations of Cognitive Grammar, Vol. 1*. Stanford University Press.

[Langacker, 2000] Langacker, R. W. (2000). A dynamic usage-based model. In Kemmer and Barlow, editors, *Topics in Cognitive Linguistics*, pages 127–161. John Benjamins, Amsterdam.

[Levelt, 1989] Levelt, P. (1989). *Speaking. From Intention to Articulation*. MIT Press.

[Litman et al., 1999a] Litman, D., Walker, M., and Kearns, M. (1999a). Automatic detection of poor speech recognition at the dialogue level. In *Proceedings of ACL'99*.

[Litman et al., 1999b] Litman, D. J., Walker, M. A., and Kearns, M. S. (1999b). Automatic detection of poor speech recognition at the dialogue level. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* College Park, Md., 20–26 June 1999, pages 309–316.

[LMF, 2005] LMF (2005). Iso-tc37/sc4-lmf language resource management: Lexical markup framework.

[Löckelt et al., 2002] Löckelt, M., Becker, T., Pfleger, N., and Alexandersson, J. (2002). Making sense of partial. In *Proceedings of the sixth workshop on the semantics and pragmatics of dialogue (EDILOG 2002)*, pages 101–107, Edinburgh, UK.

[Loetscher, 1984] Loetscher, A. (1984). Satzgliedstellung und funktionale satzperspektive. *Pragmatik in der Grammatik. Jahrbuch des Instituts fr deutsche Sprache*.

[Loos and Porzel, 2005] Loos, B. and Porzel, R. (2005). Towards ontology-based pragmatic analysis. In *In Processding of DIALOR'05*, Nancy, France.

[LuperFoy, 1992] LuperFoy, S. (1992). The representation of multimodal user interface dialogues using discourse pegs. In *Proceedings of ACL'92*.

[Maedche and Staab, 2002] Maedche, A. and Staab, S. (2002). Measuring the similarity between ontologies. In *Proceedings of the 13th Conference on Knowledge Engeneering and and Knowledge Management*, Springer, LNAI 2473, Berlin.

[MAF, 2005] MAF (2005). Iso-tc37/sc4-maf language resource management - morpho-syntactic annotation framework.

[Malaka et al., 2006] Malaka, R., Haeussler, J., Aras, H., Merdes, M., Pfisterer, D., Joest, M., and Porzel, R. (2006). Intelligent interaction with a mobile system. In Wahlster, W., editor, *SmartKom - Foundations of Multimodal Dialogue Systems*, pages 505–522. Springer, Berlin.

[Malaka and Porzel, 2000] Malaka, R. and Porzel, R. (2000). Integration of smart components for building your personal mobile guide. In *Proceedings of the ECAI Workshop on Artifical Intelligence in Mobile Systems*, Berlin, Germany.

[Malaka and Porzel, 2009] Malaka, R. and Porzel, R. (2009). Fleshing-out embodied interaction. In Mertsching, B., editor, *The 32nd Annual Conference on Artificial Intelligence*, page in press. GI.

[Malaka and Zipf, 2000] Malaka, R. and Zipf, A. (2000). Deep Map - Challenging IT Research in the Framework of a Tourist Information System. In Fesenmaier, D., Klein, S., and Buhalis, D., editors, *Information and Communication Technologies in Toursim*, pages 15–27. Springer, Heidelberg, Germany.

[Malyankar, 1999] Malyankar, R. (1999). Creating a navigation ontology. In *Workshop on Ontology Management*, Menlo Park, CA. AAAI Press.

[Mammeri, 2004] Mammeri, Z. (2004). Towards a formal model for qos specification and handling in networks. In *IWQoS*, pages 148–152.

[Markert, 1999] Markert, K. (1999). *Metonymien - Eine computerlinguistische Analyse*. Infix.

[Màrquez et al., 2008] Màrquez, L., Carreras, X., Litkowski, K. C., and Stevenson, S. (2008). Semantic role labeling: an introduction to the special issue. *Computational Linguistics*, 34(2):145–159.

[Marsh and Perzanowski, 1999] Marsh, E. and Perzanowski, D. (1999). MUC-7 evaluation of IE technology: Overview of results. In *Proceedings of the 7th Message Understanding Conference*. Morgan Kaufman Publishers.

[Masolo et al., 2002] Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., Oltramari, R., Schneider, L., Istc-cnr, L. P., and Horrocks, I. (2002). Wonderweb deliverable d17. the wonderweb library of foundational ontologies and the dolce ontology.

[Masolo et al., 2003] Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., and Schneider, L. (2003). Wonderweb deliverable d17: The wonderweb library of foundational ontologies. http://wonderweb.semanticweb.org/deliverables/documents /D18.pdf (last accessed: 04/18/2009).

[Maybury and Wahlster, 1997] Maybury, M. and Wahlster, W. (1997). *Readings in Intelligent User Interfaces*. Morgan Kaufmann.

[McCarthy, 1977] McCarthy, J. (1977). Epistemological problems of artificial intelligence. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence,* Cambridge, Mass., 22–25 August 1977, volume 2, pages 1038–1044.

[Mccarthy, 1979] Mccarthy, J. (1979). First order theories of individual concepts and propositions. In *Machine Intelligence*, pages 129–147. Edinburgh University Press.

[McCarthy, 1984] McCarthy, J. (1984). Some expert systems need common sense. *Computer Culture: The Scientific, Intellectual and Social Impact of the Computer*, 426. Annals of the New York Academy of Sciences.

[McCarthy, 1986] McCarthy, J. (1986). Notes on formalizing contexts. In Kehler, T. and Rosenschein, S., editors, *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 555–560, Los Altos, California. Morgan Kaufmann.

[McCarthy, 1990] McCarthy, J. (1990). Generality in artificial intelligence. In Lifschitz, V., editor, *Formalizing Common Sense: Papers by John McCarthy*, pages 226–236. Ablex Publishing Corporation, Norwood, New Jersey.

[McCarthy and Hayes, 1969] McCarthy, J. and Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B. and Michie, D., editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press. reprinted in McC90.

[Merriam-Webster, 2003] Merriam-Webster (2003). *Merriam Webster Collegiate Dictionary*. Merriam-Webster, Springfield, Massachusetts.

[Michaelis, 2001] Michaelis, L. A. (2001). Type shifting in construction grammar: An intergrated approach to aspectual coersion. *Cognitive Linguistics*.

[Miller et al., 1990] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

[Mögele et al., 2006] Mögele, H., Kaiser, M., and Schiel, F. (2006). Smartweb umts speech data collection the smartweb handheld corpus. In *Proceedings of the LREC Conference*, Genova, Italy.

[Mok et al., 2004] Mok, E., Bryant, J., and Feldman, J. (2004). Scaling understanding up to mental spaces. In Porzel, R., editor, *HLT-NAACL 2004 Workshop: 2nd Workshop on Scalable Natural Language Understanding*, pages 41–48, Boston, Massachusetts, USA. Association for Computational Linguistics.

[Müller, 2002] Müller, C. (2002). *Kontextabh¨ngige Bewertung der Koh¨renz von Spracherkennungshypothesen*. Master Thesis at the Institut für Informationstechnologie der Fachhochschule Mannheim.

[Müller and Strube, 2001] Müller, C. and Strube, M. (2001). MMAX: A tool for the annotation of multi-modal corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems,* Seattle, Wash., 5 August 2001, pages 45–50.

[Narayanan, 1997] Narayanan, S. (1997). *KARMA: Knowledge-Based Active Representations for Metaphor and Aspect*. PhD thesis, Computer Science Division, University of California, Berkeley, Cal.

[Narayanan, 1999a] Narayanan, S. (1999a). Moving right along: A computational model of metaphoric reasoning about events. In *Proc. Sixteenth National Conference of Artificial Intelligence (AAAI-99)*. AAAI Press, Menlo Park.

[Narayanan, 1999b] Narayanan, S. (1999b). Reasoning about actions in narrative understanding. In *Proc. Sixteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Press.

[Narayanan and Jurafsky, 1998] Narayanan, S. and Jurafsky, D. (1998). Bayesian models of human sentence processing. In *Proc. 20th Cognitive Science Society Conference*, pages 84–90. Lawrence Erlbaum Associates.

[Niles and Pease, 2001] Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In Welty, C. and Smith, B., editors, *Workshop on Ontology Management*, Ogunquit, Maine. Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001).

[Norman, 1988] Norman, D. (1988). *Psychology of Everyday Things*. Basic Books, New York.

[Nunberg, 1987] Nunberg, G. (1987). *The Pragmatics of Reference*. Indiana Linguistics University Club.

[Oberle et al., 2007] Oberle, D., Ankolekar, A., Hitzler, P., Cimiano, P., Sintek, M., Kiesel, M., Mougouie, B., Vembu, S., Baumann, S., Romanelli, M., Buitelaar, P., Engel, R., Sonntag, D., Reithinger, N., Loos, B., Porzel, R., Zorn, H.-P., Micelli, V., Schmidt, C., Weiten, M., Burkhardt, F., and Zhou, J. (2007). Dolce ergo sumo: On foundational and domain models in swinto (smartweb integrated ontology)). *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 5(3):156–174.

[Oberle et al., 2006] Oberle, D., Ankolekar, A., Hitzler, P., Cimiano, P., Sintek, M., Kiesel, M., Mougouie, B., Vembu, S., Baumann, S., Romanelli, M., Buitelaar, P., Engel, R., Sonntag, D., Reithinger, N., Loos, B., Porzel, R., Zorn, H.-P., Micelli, V., Schmidt, C., Weiten, M., Burkhardt, F., and Zhou, J. (2006). DOLCE ergo SUMO: On Foundational and Domain Models in SWIntO (SmartWeb Integrated Ontology). Technical Report AIFB, University of Karlsruhe.

[Oberle et al., 2005] Oberle, D., Lamparter, S., Eberhart, A., and Staab, S. (2005). Semantic management of web services. In *Proc. of ICSOC-2005 - 3rd Int. Conference on Service Oriented Computing, Amsterdam, The Netherlands*.

[Oerder and Ney, 1993] Oerder, M. and Ney, H. (1993). Word graphs: An efficient interface between continuous-speech recognition and language understanding. In *ICASSP Volume 2*, pages 119–122.

[Ogden and Richards, 1923] Ogden, C. K. and Richards, I. (1923). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism.* Routledge & Kegan Paul Ltd., London, UK, 10th edition.

[Pantel and Lin, 2003] Pantel, P. and Lin, D. (2003). Automatically discovering word senses. In Frederking, B. and Younger, B., editors, *HLT-NAACL 2003: Demo Session*, Edmonton, Alberta, Canada. Association for Computational Linguistics.

[Paris, 1993] Paris, C. L. (1993). *User Modeling in Text Generation.* Pinter, London.

[Pearson, 1939] Pearson, E. S. (1939). Student as a statistician. *Biometrika*, 30:210–250.

[Pereira et al., 1993] Pereira, F., Naftali, T., and Lillian, L. (1993). Distributional clustering of english words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics,* Columbus, Ohio, 22–26 June 1993.

[Perrault, 1989] Perrault, C. R. (1989). Speech acts in multimodal dialogues. In Taylor, M. M., Neel, F., and Bouwhuis, D. G., editors, *The Structure of Multimodal Dialogue*, pages 33–46. North-Holland, Amsterdam.

[Peters, 1993] Peters, K. (1993). Das textgenerierungssystem kleist im vergleich mit psycholinguistischen sprachproduktionsmodellen. die bedeutung kognitionswissenschaftlicher erkenntnisse fuer die automatische sprachgenerierung. In *17. Fachtagung fuer Kuenstliche Intelligenz*, Berlin, Germany.

[Pfleger et al., 2002] Pfleger, N., Alexandersson, J., and Becker, T. (2002). Scoring functions for overlay and their application in discourse processing. In *KONVENS-02*, Saarbrücken.

[Pfleger et al., 2003] Pfleger, N., Engel, R., and Alexandersson, J. (2003). Robust multimodal discourse processing. In *7th Workshop on the Semantics and Pragmatics of Dialogue*, Saarbrücken.

[Pinkal et al., 2000] Pinkal, M., Rupp, C. J., and Worm, K. (2000). Robust semantic processing of spoken language. In Wahlster, W., editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, Artificial Intelligence, pages 322–336. Springer, Berlin.

[Pivk et al., 2006] Pivk, A., Sure, Y., Cimiano, P., Gams, M., Rajkovic, V., and Studer, R. (2006). Transforming arbitrary tables into f-logic frames with tartar. *Data & Knowledge Engineering (DKE).*

[Poesio, 2002] Poesio, M. (2002). Scaling up anahpora resolution. In *Proceedings of the 1st Workshop on Scalable Natural Language Understanding*, pages 3–11.

[Poesio and Vieira, 1998] Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

[Ponzetto and Strube, 2006] Ponzetto, S. and Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 192–199.

[Porzel and Baudis, 2004] Porzel, R. and Baudis, M. (2004). The Tao of CHI: Towards effective human-computer interaction. In Susan Dumais, D. M. and Roukos, S., editors, *HLT-NAACL 2004: Main Proceedings*, pages 209–216, Boston, Massachusetts, USA.

[Porzel and Bryant, 2003] Porzel, R. and Bryant, J. (2003). Employing the embodied construction grammar formalism for knowledge represenation: The case of construal resolution. In *Proceedings of the 8th International Conference on Cognitive Linguistics*.

[Porzel and Gurevych, 2002] Porzel, R. and Gurevych, I. (2002). Towards context-adaptive utterance interpretation. In *Proceedings of the 3rd SIGdial Workshop,* Philadelphia, USA, 2002, pages 90–95.

[Porzel et al., 2006a] Porzel, R., Gurevych, I., and Malaka, R. (2006a). In context: Integrating domain- and situation-specific knowledge. In Wahlster, W., editor, *SmartKom - Foundations of Multimodal Dialogue Systems*, pages 269–284. Springer Verlag.

[Porzel et al., 2003a] Porzel, R., Gurevych, I., and Müller, C. (2003a). Ontology-based contextual coherence scoring. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue,* Saporo, Japan, July 2003.

[Porzel and Malaka, 2004a] Porzel, R. and Malaka, R. (2004a). A task-based approach for ontology evaluation. In *ECAI-2004 Workshop on Ontology Learning and Population*, Valencia, Spain.

[Porzel and Malaka, 2004b] Porzel, R. and Malaka, R. (2004b). Towards measuring scalability in natural language understanding tasks. In Porzel, R., editor, *HLT-NAACL 2004 Workshop: 2nd Workshop on Scalable Natural Language Understanding*, pages 9–16, Boston, Massachusetts, USA. Association for Computational Linguistics.

[Porzel and Malaka, 2005] Porzel, R. and Malaka, R. (2005). A task-based framework for ontology learning, population and evaluation. In Buitelaar, P., Cimiano, P., and Magnini, B., editors, *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123 of *Frontiers in Artificial Intelligence*, pages 107–122. IOS Press.

[Porzel et al., 2006b] Porzel, R., Micelli, V., and Aras, H. (2006b). Tying the knot: Ground entities, descriptions and information objects for construction-based information extraction. In *In Proceedings of Ontolex 2006*.

[Porzel et al., 2003b] Porzel, R., Pfleger, N., Merten, S., Löckelt, M., Engel, R., Gurevych, I., and Alexandersson, J. (2003b). More on Less: Further Applications of Ontologies in Multi-Modal Dialogue Systems. In *Proceedings of the 3rd IJCAI 2003 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Acapulco, Mexico.

[Porzel and Strube, 2002] Porzel, R. and Strube, M. (2002). Towards context-dependent natural language processing. In Klenner, M. and Visser, H., editors, *Computational Linguistics for the New Millennium: Divergence or Synergy*. Peter Lang Academic Publishers, Berlin.

[Porzel et al., 2006c] Porzel, R., Zorn, H.-P., Loos, B., and Malaka, R. (2006c). Towards a seperation of pragmatic knowledge and contextual information. In *ECAI'06 Workshop on Context and Ontology, Riva del Garda, August 28,*.

[Puglisi et al., 2008] Puglisi, A., Baronchelli, A., and Loreto, V. (2008). Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences*, pages 0802485105+.

[Ramsey, 2000] Ramsey, A. (2000). Speech act theory and epistemic planning. In Bunt, H. and Black, W., editors, *Computational Pragmatics, Abduction, Belief and Context; Studies in Computational Pragmatics*, pages 293–310. John Benjamins, Amsterdam.

[Rapp and Strube, 2002] Rapp, S. and Strube, M. (2002). An iterative data collection approach for multimodal dialogue systems. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

[Rapp et al., 2000] Rapp, S., Torge, S., Goronzy, S., and Kompe, R. (2000). Dynamic speech interfaces. In *Proceedings of 14th ECAI WS-AIMS*.

[Rast, 2007] Rast, E. H. (2007). *Reference and Indexicality*. Number 17 in Logische Philosophie. Logos Verlag.

[RDF, 2001] RDF (2001). http://www.w3.org/rdf/. (last accessed: 10/10/2008).

[RDFS, 2001] RDFS (2001). http://www.w3.org/tr/rdf-schema/. (last accessed: 10/10/2008).

[Reithinger et al., 2003] Reithinger, N., Alexandersson, J., Becker, T., Blocher, A., Engel, R., Löckelt, M., Müller, J., Pfleger, N., Poller, P., Streit, M., and Tschernomas, V. (2003). Smartkom: adaptive and flexible multimodal access to multiple applications. In *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*, pages 101–108, New York, NY, USA. ACM.

[Reithinger et al., 2005] Reithinger, N., Bergweiler, S., Engel, R., Herzog, G., Pfleger, N., Romanelli, M., and Sonntag, D. (2005). Look under the hood: Design and development of the first smartweb system demonstrator. In *Proceedings of IMCI*, Trento, Italy.

[Reithinger and Maier, 1995] Reithinger, N. and Maier, E. (1995). Utilizing statistical speech act processing in Verbmobil. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* Cambridge, Mass., 26–30 June 1995, pages 116–121.

[Rosario et al., 2002] Rosario, B., Hearst, M., and Fillmore, C. (2002). The descent of hierarchy, and selection in relational semantics. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 247–254.

[Rosch, 1983] Rosch, E. (1983). Prototype classification and logical classification: The two systems. *Scholnick, E.K. (Hrsg.): New Trends in Conceptual Representation: Challenges to Piaget's Theory?*, pages 73–86.

[Rosenfeld and Feldman, 2006] Rosenfeld, B. and Feldman, R. (2006). Ures: an unsupervised web relation extraction system. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 667–674, Morristown, NJ, USA. Association for Computational Linguistics.

[Rüggenmann and Gurevych, 2004a] Rüggenmann, K. and Gurevych, I. (2004a). Assigning domains to speech recognition hypotheses. In *Proceedings of HLT-NAACL Workshop on Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Knowledge for Speech Processing*, pages 70–77, Boston, USA.

[Rüggenmann and Gurevych, 2004b] Rüggenmann, K. and Gurevych, I. (2004b). Assigning domains to speech recognition hypotheses. In *Proceedings of HLT-NAACL Workshop on Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Knowledge for Speech Processing*, pages 70–77, Boston, USA.

[Russell and Norvig, 1995] Russell, S. J. and Norvig, P. (1995). *Artificial Intelligence. A Modern Approach*. Prentice Hall, Englewood Cliffs, N.J.

[Sack et al., 1974] Sack, S., Schegloff, E., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50.

[Schiel et al., 2002] Schiel, F., Steininger, S., and Türk, U. (2002). The SmartKom Multimodal Corpus at BAS. In *Proceedings of the 3rd LREC*, Las Palmas Spain.

[Schiel and Türk, 2006] Schiel, F. and Türk, U. (2006). Wizard-of-oz recordings. In Wahlster, W., editor, *SmartKom - Foundations of Multimodal Dialogue Systems*, pages 571–598. Springer Verlag.

[Schilit et al., 1994] Schilit, B., Adams, N., and Want, R. (1994). Context-aware computing applications. In *In Proceedings of the Workshop on Mobile Computing Systems and Applications*, pages 85–90. IEEE Computer Society.

[Schober, 1993] Schober, M. (1993). Spatial perspective-taking in conversation. *Cognition*, 47(1):1–23.

[Schütze, 1998] Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–124.

[Schwartz and Chow, 1990] Schwartz, R. and Chow, Y. (1990). The n-best algorithm: an efficient and exact procedure for finding the n most likely sentence hypotheses. In *Proceedings of ICASSP'90, Albuquerque, USA*.

[Searle, 1975] Searle, J. R. (1975). Indirect speech acts. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 59–82. Academic Press, San Diego, CA.

[Sebanz et al., 2006] Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, 10:70–76.

[Shankar et al., 2000] Shankar, T. R., VanKleek, M., Vicente, A., and Smith, B. K. (2000). A computer mediated conversational system that supports turn negotiation. In *Proceedings of the Hawai'i International Conference on System Sciences*, Maui, Hawaii.

[Shen and Lapata, 2007] Shen, D. and Lapata, M. (2007). Using semantic roles to improve question answering in proceedings emnlp-conll, 2007. In *Proceedings EMNLP-CoNLL*.

[Shepard, 1975] Shepard, R. (1975). *Form, Formation, and Transformation of Internal Representations*. Erlbaum, Hillsdale, N.J.

[Shepard and Metzler, 1971] Shepard, R. and Metzler, J. (1971). Mental rotation of three dimensional objects. *Science*, 171(3):701–703.

[Shneiderman and Plaisant, 2004] Shneiderman, B. and Plaisant, C. (2004). *Designing the User Interface : Strategies for Effective Human-Computer Interaction (4th Edition)*, chapter 12.3: Nonanthropomorphic Design, pages 484–490. Addison Wesley.

[Shriberg et al., 2004] Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The icsi meeting recorder dialog act (mrda) corpus. In Strube, M. and Sidner, C., editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

[Soanes and Stevenson, 2005] Soanes, C. and Stevenson, A. (2005). *Oxford Dictionary of English*. Oxford University Press, Oxford, UK.

[Sonntag et al., 2007] Sonntag, D., Engel, R., Herzog, G., Pfalzgraf, A., Pfleger, N., Romanelli, M., and Reithinger, N. (2007). SmartWeb Handheld - Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services. In Huang, T. S., Nijholt, A., Pantic, M., and Pentland, A., editors, *Artifical Intelligence for Human Computing*, volume 4451 of *Lecture Notes in Computer Science*, pages 272–295. Springer, Heidelberg, Germany.

[Steels, 1998a] Steels, L. (1998a). The origins of ontologies and communication conventions in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 1:169–194.

[Steels, 1998b] Steels, L. (1998b). Spontaneous lexicon change. In *In Proceedings of COLING-ACL*, pages 1243–1249. ACL.

[Steels, 2001] Steels, L. (2001). Language games for autonomous robots. *IEEE Intelligent systems*, pages 16–22.

[Steels, 2008] Steels, L. (2008). The symbol grounding problem has been solved. so what's next? In de Vega, M., editor, *Symbols and Embodiment: Debates on Meaning and Cognition*, chapter 12. Oxford University Press, Oxford.

[Stent et al., 1999] Stent, A., Dowding, J., Gawron, J. M., Owen Bratt, E., and Moore, R. (1999). The Command Talk spoken dialogue system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* College Park, Md., 20–26 June 1999, pages 183–190.

[Stevenson, 2002a] Stevenson, M. (2002a). *Combining disambiguation techniques to enrich an ontology.* In Proceedings of the 15th ECAI workshop on Machine Learning and Natural Language Processing for Ontology Engineering.

[Stevenson, 2002b] Stevenson, M. (2002b). *Word Sense Diambiguation.* CSLI Publications.

[Stevenson, 2003] Stevenson, M. (2003). *Word Sense Disambiguation: The Case for Combining Knowldge Sources.* CSLI.

[Stevenson and Wilks, 2001] Stevenson, M. and Wilks, Y. (2001). The Interaction of Knowledge Sources in Word Sense disambiguation. *Computational Linguistics*, 27(3):321–349.

[Strube and Müller, 2003] Strube, M. and Müller, C. (2003). A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 168–175, Sapporo, Japan. Association for Computational Linguistics.

[Strube and Wolters, 2000] Strube, M. and Wolters, M. (2000). A probabilistic genre-independent model of pronominalization. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics,* Seattle, Wash., 29 April – 3 May, 2000, pages 18–25.

[Sussna, 1993] Sussna, M. (1993). Word sense disambiguation for free text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management*.

[Sutton and Barto, 1998] Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press.

[Svb-Zamazal et al., 2008] Svb-Zamazal, O., Svtek, V., Meilicke, C., and Stuckenschmidt, H. (2008). Testing the impact of pattern-based ontology refactoring on ontology matching results. In Shvaiko, P., Euzenat, J., Giunchiglia, F., and Stuckenschmidt, H., editors, *CEUR Workshop*, volume 431. CEUR-WS.org.

[Sweetser, 2003] Sweetser, E. (2003). Levels of meaning in speech and gesture: Real space mapped onto epistemic and speech-interactional mental spaces. In *Proceedings of the 8th International Conference on Cognitive Linguistics*, Logrono, Spain.

[SynAF, 2005] SynAF (2005). Iso-tc37/sc4-synaf language resource management - syntactic annotation framework.

[Takhtamysheva, 2009] Takhtamysheva, A. (2009). *Collecting Language Data through Games*. Master Thesis at the University of Applied Sciences Bremerhaven.

[Talmy, 1988] Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, 12:49–100.

[Tomasello, 2008] Tomasello, M. (2008). *The Origins of Human Communication*. MIT Press.

[Turner, 1993] Turner, E. H. (1993). Exploiting predictions to organize goals in real-world domains. In *AAAI Spring Symposium on Foundations of Automatic Planning: The Classical Approach & Beyond*, pages 142–145, Stanford, CA.

[Van Rijsbergen, 1979] Van Rijsbergen, C. J. (1979). *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow.

[von Stutterheim and Klein, 1989] von Stutterheim, C. and Klein, W. (1989). Referential movement in descriptive and narrative discourse. In Dietrich, R. and Graumann, C. F., editors, *Language Processing in Social Context*, pages 39–76. Elsevier Science Publisher, Amsterdam, The Netherlands.

[Voorhees, 1993] Voorhees, E. M. (1993). Using wordnet to disambiguate word senses for text retrieval. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 171–180, New York, NY, USA. ACM Press.

[W3C-OEP, 2005] W3C-OEP (2005). Semantic web best practices and deployment working group: Ontology engineering and patterns task force. http://www.w3.org/2001/sw/BestPractices/OEP/ (last accessed: 04/18/2009).

[Wahlster, 2003] Wahlster, W. (2003). Smartkom: Symmetric multimodality in an adaptive and reusable dialog shell. In DLR, editor, *Proceedings of the of the 26th German Conference on Artificial Intelligence, September 2003, Hamburg, Germany*.

[Wahlster, 2004] Wahlster, W. (2004). Smartweb: Mobile applications of the semantic web. In *Proceedings of Informatik 2004*.

[Wahlster et al., 1993] Wahlster, W., Andre, E., Finkler, W., Profitlich, H. J., and Rist, T. (1993). Plan-based integration of natural language and graphics generaion. *Artificial Intelligence*, 63:387–427.

[Wahlster et al., 2001] Wahlster, W., Reithinger, N., and Blocher, A. (2001). SmartKom: Multimodal Communication with a Life-like Character. In *Procedings of Eurospeech*, pages 1547 – 1550, Aalborg, Denmark.

[Walker et al., 2000] Walker, M. A., Kamm, C. A., and Litman, D. J. (2000). Towards developing general model of usability with PARADISE. *Natural Language Engeneering*, 6.

[Wallis and Shortliffe, 1982] Wallis, J. and Shortliffe, E. (1982). Explanatory power for medical expert systems. *Methods of Information in Medicine*, 6(3):127–136.

[Waltz, 1978] Waltz, D. (1978). An english language question answering system for a large relational database. *Communications of the ACM*, 21(7).

[Wazinski, 1992] Wazinski, P. (1992). Generating spatial descriptions for cross-modal references. In *Proceedings of the third conference on Applied natural language processing*, pages 56–63, Morristown, NJ, USA. Association for Computational Linguistics.

[Webber, 1991] Webber, B. (1991). Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

[Webber, 1979] Webber, B. L. (1979). *A formal approach to discourse anaphora*. Garland, New York, N.Y.

[Weinhammer and Rabold, 2003] Weinhammer, K. and Rabold, S. (2003). Durational Aspects in Turn Taking. In *Proceedings of International Conference Phonetic Sciences*, Barcelona, Spain.

[Weiss, 1973] Weiss, S. (1973). Learning to disambiguate. *Information Storage and Retrieval*, 9.

[Whitehead, 1898] Whitehead, A. N. (1898). *A Treatise on Universal Algebra with Applications*. Cambridge Universitity Press (Reprint 1960), Cambridge, UK.

[Widdows, 2003a] Widdows, D. (2003a). A Mathematical Model for Context and Word-Meaning. In *Fourth International and Interdisciplinary Conference on Modeling and Using Context*, Stanford, California.

[Widdows, 2003b] Widdows, D. (2003b). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *HLT-NAACL 2003: Main Proceedings*, pages 276–283, Edmonton, Canada. ACL.

[Wilks, 2002] Wilks, Y. (2002). Ontotherapy: or how to stop worrying about what there is. In *Proceedings of the Workshop on Ontologies and Lexical Knowledge Bases*, Las Palmas, Canary Islands.

[Wilson, 1997] Wilson, M. (1997). Metaphor to Personality: the role of animation in intelligent interface agents. In *Proceedings of the IJCAI-97 Workshop on Animated Interface Agents: Making them Intelligent*.

[Winograd, 1972] Winograd, T. (1972). *Understanding Natural Language*. Academic Press, New York.

[Winston, 1992] Winston, P. H. (1992). *Artificial Intelligence*. Addison-Wesley.

[WonderWeb, 2003] WonderWeb (2003). Ontology infrastructure for the semantic web. http://wonderweb.semanticweb.org/index.shtml (last accessed: 04/18/2009).

[Woodburn et al., 1991] Woodburn, R., Procter, R., Arnott, J., and Newell, A. (1991). A study of conversational turn-taking in a communication aid for the disabled. In *People and Computers*, pages 359–371. Cambridge University Press, Cambridge.

[Woods, 1977] Woods, W. (1977). Lunar rocks in natural english: Explorations in natural language question answering. In Zampoli, A., editor, *Lingusitic Structures Processing*, pages 521–569. Elsevier North-Holland, New York.

[Wooffitt et al., 1997] Wooffitt, R., Gilbert, N., Fraser, N., and McGlashan, S. (1997). *Humans, Computers and Wizards: Conversation Analysis and Human (Simulated) Computer Interaction.* Brunner-Routledge, London.

[WSDL, 2001] WSDL (2001). http://www.w3.org/tr/wsdl20/. (last accessed: 04/19/2009).

[XML, 2001] XML (2001). http://www.w3.org/xml/. (last accessed: 10/10/2008).

[Yarowsky, 1992] Yarowsky, D. (1992). Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the 15th International Conference on Computational Linguistics,* Nantes, France, 23-28 August 1992, volume 1.

[Yarowsky, 1995] Yarowsky, D. (1995). Unsupervised word sense disambiguation rivalling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* Cambridge, Mass., 26–30 June 1995, pages 189–196.

[Yngve, 1970] Yngve, V. (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, Chicago, Illinois.

[Zhou et al., 2007] Zhou, G., Zhang, M., Ji, D., and Zhu, Q. (2007). Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 728–736, Prague, Czech Republic. Association for Computational Linguistics.

[Zoeppritz, 1985] Zoeppritz, M. (1985). Computer talk? Technical report, IBM Scientific Center Heidelberg Technical Report 85.05.

[Zwaan and Radvansky, 1998] Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123:162–185.