Dialogue Acts: One or More Dimensions?

Andrei Popescu-Belis

ISSCO Working paper n. 62

ISSCO
School of Translation and Interpretation
University of Geneva
Bd. du Pont-d'Arve 40
1211 Geneva 4, Switzerland
andrei.popescu-belis@issco.unige.ch

November 2005 (updated August 2007)

Abstract

This report surveys the main theories of dialogue and communication that have been used to devise dialogue act tagsets, distinguishing theories that deal with a specific level of communication from theories that integrate several levels. The report proceeds to analyse four dialogue act tagsets that have been used to annotate large scale dialogue corpora (DAMSL, SWBD-DAMSL, ICSI-MRDA and MALTUS), paying particular attention to the difference between the range of possible combinations of tags and the range of combinations that do occur in the hand-labelled data.

The problem of the dimensionality of tagsets is then introduced, and discussed in relation with other factors that influence the performance of automatic dialogue act taggers. After a brief discussion of empirical arguments relevant to the dimensionality problem, derived from human and automatic labelling experiments, the report proposes a synthesis of guidelines for the definition of dialogue act tagsets, based on multi-dimensional theoretical inspiration, and cross-dimensional constraints and the notion of dominant utterance-function, in order to reduce the search for automatic dialogue act taggers.

As a compromise between the divergent needs for theoretical grounding and tagging accuracy, the Dominant Function Approximation proposes that automatic dialogue act taggers could focus initially on finding the main dialogue function of each utterance. This approximation is shown to be empirically acceptable and to have significant practical relevance.

1 Introduction

The functions of utterances in dialogue-based interaction are often referred to, collectively, as *dialogue acts* (henceforth DA), especially in the dialogue models used by computational linguists. The automatic recognition of DAs plays an essential role in human-computer dialogue systems (Traum, 1999; Sadek, 2000), but also in human dialogue understanding applications (Zechner, 2002), spoken language translation (Reithinger & Maier, 1995), or automatic speech recognition (Stolcke *et al.* , 2000). Automatic DA tagging has even acquired an interest of its own, and new techniques for increasing the performance of DA taggers have been proposed. However, DA taggers that use different tagsets are often difficult to compare, while dialogue corpora annotated using a given tagset cannot be easily reused in projects with a different tagset. The choice of a specific DA tagset has a also significant impact on the accuracy of automatic tagging.

In this report, we use theoretical as well as empirical arguments drawn from several tagsets that have been used to label extensive corpora in order to answer one of the main questions that influence the choice of a DA tagset: are one-dimensional tagsets preferable to multi-dimensional ones, or vice-versa? The answer depends on many considerations, among which we will analyse here the role of dialogue theories and the various techniques used to maximize the accuracy of automatic DA taggers, which is the main application envisioned here. As taggers often use statistical machine learning, the notions of search space and feature space will play a crucial role in the discussion.

We will argue that, while theoretical inspiration and ease of annotation favour multi-dimensional tagsets, there are computational reasons to prefer one-dimensional tagsets. We suggest that a better integration between theory and practice should lead to the design of constrained multi-dimensional tagsets, which could remain computationally efficient while benefiting from a sounder theoretical basis. The definition of such tagsets requires strong cross-dimensional constraints derived from theories of dialogue, as well as statistical learning methods that can use such constraints to train multi-dimensional classifiers. A constraint that increases even more computational efficiency is also proposed, based on the hypothesis that utterances have a single dominant function, while functions in other dimensions have, implicitely, a default value. Dialogue act tagging would then aim at finding the dominant function of each utterance, a task of significant practical value for language engineering applications.

We summarize below the main constraints on the definition of a DA tagset (Section 2), in particular the application and the dialogue genre. Then we review several theories regarding the function of utterances in dialogues (Section 3) and introduce four tagsets that have been used to annotate large scale resources: : DAMSL, SWBD-DAMSL, ICSI-MRDA, and

MALTUS (Section 4). From a computational point of view, we distinguish tagsets with mutually-exclusive tags (exactly one tag per utterance) from multi-dimensional tagsets (one or more tags per utterance), first theoretically (Section 5), then in relation to automatic DA tagging (Section 6). Empirical data from annotated corpora and human/automatic tagging experiments shows that dimensions of DAs that originate in independent theories appear to be related in reality (Section 7). Based on these observations, we finally summarize (Section 8) our main argument for multi-dimensional theoretical grounding combined with cross-dimensional constraints, based on the distinction of dominant and default utterance-functions.

2 Definition(s) of Dialogue Acts

2.1 What is a Dialogue Act?

When language is used in interactions between individuals, the utterances that these individuals exchange generally modify their cognitive status and the state of the surrounding part of the world, i.e. the dialogue context. The effect of an utterance on the context is often called a *dialogue act*. For instance, if speaker A would like speaker B to close the window of the room where they are located, A could use various utterances to achieve this goal, and these utterances would all share the same dialogue function, i.e. fulfil the same dialogue act. The definition of the DA concept is often part of a specific theory of dialogue. According to one of the most general definitions, proposed by Bunt (2000, page 144), a DA is "the combination of a communicative function and a semantic content", where the former is "the way in which dialogue participants use information to change the context."

The occurrence of a DA (or token) must be distinguished from its category or type of DA. For instance, the question "Can you pass the salt?" is an occurrence of the more general type 'question'. In practice, the term 'dialogue act' can be used to refer to a single occurrence or to a type (just as 'word' can be used for both tokens and types, as in "The abstract has 100 words" vs. "I learned a new word today"). The term 'dialogue function' refers less ambiguously to the type of DA.

Speech acts are a well-known aspect of communicative function (see Section 3.3 below), but, especially in computational or language engineering approaches, the notion of DA covers a larger range of functions, and include speech acts: "while speech acts provide a useful characterization of one kind of pragmatic force, more recent work, especially in building dialogue systems, has significantly expanded this core notion, modelling more kinds of conversational functions that an utterance can play. The resulting enriched acts are called dialogue acts or conversational moves" (Jurafsky & Martin,

2000, chapter 19.2, page 929)[1]. In their multi-level approach to *conversation acts*, Traum and Hinkelman (1992) distinguish "four levels of action necessary for maintaining the coherence and content of conversation." The four levels of conversation acts are: turn-taking acts, grounding acts, core speech acts, and argumentation acts.

Many studies in computational linguistics or language engineering assume a definition of the notion of DA along the previous lines, and proceed to define types of DAs according to their particular goals. In other words, many studies define *DA tagsets* that enumerate the possible dialogue functions they consider, depending on a variety of factors that we examine in the following section (2.2). A *DA tag* stands for a particular dialogue function that an utterance can fulfil. *Tagging* an utterance means determining one or more dialogue functions fulfilled by that utterance, and assigning it one or more DA tags. The *label* of the utterance denotes the set of tags attributed to that utterance, but the the distinction between tags and labels does not seem to be universally accepted. Following recent examples (Dhillon *et al.* , 2004) we will nevertheless distinguish these terms throughout the paper.

2.2 Constraints on the Definition of DA Tagsets

The objective of the study has presumably the most important influence on the choice of a DA tagset. Some studies investigate the mechanisms of human dialogue, while (many) others envisage task-oriented human-computer dialogue systems. Other applications combine these two goals, such as the modelling of human-human dialogue for the design and evaluation of automatic customer service applications (Rosset & Lamel, 2004), or the analysis of human-computer dialogues to automatically evaluate their quality in terms of user satisfaction (Walker & Passonneau, 2001). Or, for instance, SWBD-DAMSL (Jurafsky *et al.* , 1997) was designed to provide an extra feature to improve speech recognition on free two-party conversations (Stolcke *et al.* , 2000). The Verbmobil tagset was designed to help the automatic translation of two-party spoken dialogue about appointment scheduling (Jekat *et al.* , 1995). A recent overview of the various purposes of DA tagsets was proposed by Bunt, who distinguishes (1) conceptual analyses of human dialogue, (2) dialogue systems, (3) corpus annotation for empirical analyses, and (4) agent communication protocols (Bunt, 2005, pages 2–3).

Two other important distinctions are related to the intended application of a DA tagset: the dialogue domain and the activity type. In an overview of existing tagsets—oriented towards language engineering applications—the following range of observed domain restrictions was identified: travel, transport, computer operating systems, courtroom interaction, business ap-

[1] An updated version of this chapter on 'Dialogue and Conversational Agents' is available at: `http://www.cs.colorado.edu/~martin/SLP/updated.html` (draft of May 18, 2005).

pointments, directory enquiry services, furnishing rooms, giving directions / instructions (Klein & Soria, 1998). The same report identifies also the following activity types: cooperative negotiation, information extraction, problem solving, teaching/instruction, counselling, chatting. Tagsets defined for different activity types may differ quite significantly, since not all types of DAs are present in all types of interaction. Moreover, apparently similar types of DAs can have very different functions according to the activity type, as shown by Levinson (1992) in the case of questions. A useful typology of interactions was proposed by McGrath for small group meetings, using the 'circumplex model' (1984)[2].

The "20 questions on DA taxonomies" proposed by Traum (2000) suggest further, finer-grained criteria for the design of a DA tagset. The first 15 questions pertain to the theoretical grounding of the types of DAs considered in a tagset, for instance how speakers' intentions are taken into account in the tagset. Questions 16, 17 and 20 consider the empirical validation and the intended application of the tagset. Questions 18 and 19 target the required complexity of the tagset.

We proposed elsewhere a summary of the criteria used to define a DA tagset for automatic tagging, in six classes (Popescu-Belis, 2003, page 3). These constraints were based on our experience with the definition of the MALTUS tagset used for the analysis of multi-party meetings. The tagset should first be related to a theory of dialogue (1), and should be compatible with observed functions of actual utterances, i.e. adapted to the dialogue genre (2). It should also be empirically validated by high inter-annotator agreement figures (3). The possibility of automatic recognition of the DAs defined from the tagset should be considered (4), depending on the particular NLP application in mind (5). Some form of mapping to existing tagsets (6) helps reusing previously annotated data and ensures that valuable insights from previous work are preserved.

In this article, we focus on DA tagsets designed for automatic DA tagging, without specifying an application or a domain any further. If the design purpose was radically different, e.g. to elicit pragmatic judgments from human annotators, then the design principles for a DA tagset would be quite diffrent. Our main concern in the context of automatic DA tagging is: are mono-dimensional DA tagsets to be preferred over multi-dimensional ones, or not?

Traum (2000) depicts the two extremes, mono- and multi-dimensional tagsets:

> Given that utterances in dialogue are generally multi-functional,

[2]MCGrath defined eight classes of dialogues, using two main oppositions: cooperation vs. conflict and behavioural vs. conceptual, and using four 'quadrants': generate, choose, negotiate, execute. The eight resulting types of interaction are: planning, creativity, intellective, decision-making, cognitive-conflict, mixed-motive, contest/competition, and performance/psycho-motor (McGrath, 1984, pages 60–66).

> the question arises as to how best to capture this multiplicity of functions in a taxonomy. There are two extremes: one is to separate out each function and code it separately, which requires multiple labels for each utterance, one for each function [...] The other extreme is to combine sets of coherent bundles of dialogue functions into complex labels [...] It is also possible to find taxonomies that take a more intermediate position... (Traum, 2000, page 23)

These options—exemplified respectively by the DAMSL, SWBD-DAMSL and Verbmobil tagsets—are the object of the present discussion.

3 Dialogue Acts in Linguistics and Pragmatics

Modelling dialogue-based interaction between individuals involves analyses of the linguistic content and, more generally, of the situation in which communication takes place. Given the complexity of the phenomenon, various theoretical approaches to dialogue have been proposed. Each approach is often concerned more specifically with one aspect of dialogue, and proceeds to study it within the paradigm of an existing discipline.

We review below the main contributions to the study of dialogue from linguistics and from pragmatics. These contributions are often linked to a particular level of organisation of dialogue-based interactions, i.e. to a particular granularity of the subdivision of dialogues in more elementary parts. DAs are often associated—in existing theories—to utterances; therefore, we first discuss this important unit of dialogue.

3.1 Utterances and DAs

It has long been acknowledged that the utterance is a main unit of spoken language, which somewhat parallels sentences in written language, with a series of important differences (Lyons, 1981, section 5.5). For instance, utterances are often incomplete, or include disfluencies; their syntax is often simpler than that of written sentences; and they contain more fillers or discourse markers their written counterparts (Brown & Yule, 1983, section 1.2.6, pages 14–19). Therefore, the identification of utterances in spoken discourse is not only based on the syntactic unity of its constituting words, but also on prosodic factors such as silences, pitch contour and stress (Jurafsky & Martin, 2000, section 9.9)[3].

[3]Utterances have sometimes been defined as the "stretch of talk, by one person, before and after which there is silence on the part of that person" (Harris, 1951, page 14). An utterance "may consist of a single word, a single phrase or a single sentence [...], a sequence of sentences [...], one or more grammatically incomplete sentence-fragments; [...] there is no simple relation of correspondence between utterances and sentences" (Lyons,

Utterances can be studied from a lexical, syntactic or semantic point of view, but the analysis of their *function* in dialogue-based interaction pertains to pragmatics. Apart from being samples of linguistic performance, utterances are also functional units of dialogue, with specific roles as DAs in the various theories outlined below.

In addition, sub-utterance and supra-utterance units of dialogue structure have also been proposed. For instance, in Traum and Hinkelman's (1992) multi-level approach (see section 3.9), functions may appear at four levels: sub-utterance, utterance, discourse unit, and multi-utterance units. In other theories, one or more contiguous utterances constitute a dialogue move, and one or more moves constitute a turn (Stenström, 1994, chapter 2). A hierarchical organization of dialogue is also hypothesized by the dialogue grammar theory of the "Geneva school" (Roulet *et al.* , 1985; Moeschler, 1989).

3.2 Sentence-types vs. utterance-functions

According to Levinson (1983, page 242), the function of an utterance must be clearly separated from its sentence-type. In English, the three basic sentence-types, or linguistic forms of sentences, are the imperative, interrogative, and declarative. The sentence-type is somewhat related to the mood of the verb phrase in the main clause (indicative, imperative, subjunctive, conditional), but also to other lexical and syntactic markers. The notion of sentence type appears to be a linguistic universal (Sadock & Zwicky, 1985).

Sentence-types seem correlated with certain utterance-functions, most obviously imperatives with orders, and interrogatives with questions. A major challenge of pragmatics is to show how the sentence-type, the other properties of the utterance, and the contextual factors contribute to determine the utterance-function. Indeed, some DAs "can be signalled directly by surface features of discourse, although usually a combination of surface features and context will be necessary to disambiguate acts" (Traum & Hinkelman, 1992, page 79). For instance, the function of utterances with the sentence-type 'question' varies considerably according to the setting in which they are used, e.g., in exams, courts, greetings, lecture openings, etc. (Levinson, 1992). The analysis of such an utterance as a request for information, as in speech act theory, does not hold universally.

The main aspects of utterance-function that we will consider here are all inspired by major, albeit separate, theoretical contributions to linguis-

---

1977, pages 26-27).

However, quantifying the amount of silence required to delimit an utterance is a problematic issue: "the pause must not be filled by another communicator's contribution, nor must it be so long that it is more reasonable to regard renewed activation as a new contribution" (Allwood, 2000, section 6). Therefore, the above definition seems to better characterize *speaker turns* rather than utterances, which are best conceived as parts of a turn that accomplish an elementary dialogue function.

tics, discourse studies and pragmatics. We will first outline a number of dimensions that have been studied separately (3.3–3.7), and then analyze some multi-dimensional, integrative approaches (3.9). Most of them have shaped, to a variable extent, the four DA tagsets that we analyze later on in this paper: speech acts (3.3), turn-taking (3.4), adjacency pairs (3.5), topic organization (3.6), politeness (3.8), and rhetorical structure (3.7).

At the time the theory of speech acts was being developed, a taxonomy of communicative acts was already proposed by Bales (1950) in his Interaction Processes Analysis (IPA). The twelve types of interaction units pertain either to the task of the group, or to the social-emotional area (positive or negative). In the task area, types of utterances that ask for orientations/opinions/suggestions are matched by types of utterances that give orientations/opinions/suggestions[4]. These categories are reflected in the more recent studies of speech acts, adjacency pairs, and politeness phenomena. Variants of the IPA scheme are still used for dialogue annotation (Carletta & Kilgour, 2005).

3.3 Speech acts

According to Austin (1962), Searle (1969), and many others, when speakers utter certain words, they also perform certain actions by virtue of pronouncing them. Speech act theory provides a taxonomy of such actions, that is, of the illocutionary force of utterances, which is distinct from their semantic content. Semantic content can be true or false, but a speech act can only be felicitous or infelicitous. The types of illocutionary force, often used to define DA tagsets, are formally defined in terms of logic-based pre- and post-conditions on the speakers' mental states (Searle, 1976; Vandervecken, 1990). (1) *Representatives* commit the speaker to the truth of the expressed proposition: assertion, conclusion, etc. (2) *Directives* are attempts by the speaker to get the addressee to do something: request, question, suggestion, etc. (3) *Commissives* commit the speaker to a future course of action: promise, threat, offer, etc. (4) *Expressives* express a psychological state: thanks, apologize, welcome, congratulation, etc. (5) *Declarations* operate immediate changes in the institutional state of affairs: christening, firing from employment, excommunication, declaration of war, etc.

One of the main difficulties in determining the illocutionary force of an utterance is the problem of indirect speech acts, e.g. (Levinson, 1983, pages 263–273). Beyond its sentence-type and the direct speech act it conveys (i.e. its literal illocutionary force), an utterance may indirectly convey another speech act, which constitutes its indirect illocutionary force, and is

[4]'Orientation' covers also information, repetition, confirmation; 'opinion' covers also evaluation, analysis, expression of feeling; 'suggestion' covers also direction, possible ways of action. These elements evoke the first speech-act proposals, of which they are contemporaneous.

often the dominant function. For instance, when uttered at the table, "Can you pass the salt?" is an interrogative, its literal force is a question (apparently a demand for information), but its indirect and dominant force is a request, namely a conventional demand to pass the salt. Uttering only in response "Yes, I can" is not enough, as long as the request is not complied with or cancelled. Many solutions have been proposed to the problem of indirect speech acts, such as the theory of generalized conversational implicatures (Levinson, 2000) or a discourse-semantic account based on rhetorical relations (Asher & Lascarides, 2001), but none of them seems to be entirely satisfactory (Lycan, 2000, pages 199–202).

These difficulties, however, do not question the taxonomy of speech acts, which is a frequent starting point for language processing applications (Traum, 1999; Sadek, 2000; Jurafsky, 2003). In fact, in "literal" dialogues, ignoring for instance politeness phenomena, indirect illocutionary force may often be absent. Literal force can be computed in this case either from an explicit performative verb, or using the fact that "the three major sentence-types in English, namely the imperative, interrogative and declarative, have the forces traditionally associated with them, namely ordering (or requesting), questioning and stating respectively" (Levinson, 1983, page 263).

While many studies assume that speech acts are realized by individual utterances, several researchers argued that a speech act must be *grounded* or acknowledged by the hearer in order to count as a realized speech act and to become common ground in the dialogue (Jurafsky & Martin, 2000, chapter 19, page 724 sq.). The analysis of speech acts should thus be accompanied by an account of *joint* linguistic acts accomplished by speakers and hearers, as proposed by H. H. Clark (Clark & Schaefer, 1989; Clark, 1996). Joint acts consist of a presentation phase and an acceptance phase. Following this line of thought, speech acts pertain, strictly speaking, to discourse units larger than utterances, which ground the speech acts through successive utterances produced by the speaker and the hearer (Traum & Hinkelman, 1992, section 2.1), while individual utterances play specific, grounding-related roles. However, according to Traum and Hinkelman, the "initial presentation" of a speech act in an utterance "is best considered as a speech act attempt, which is not fully realized until its DU is grounded." This allows the simplification of associating a speech act to an utterance only when there is no misunderstanding of the speech act presentation.

3.4 Turn management

Findings from conversation analysis show that some utterances, or fragments of utterances, serve more particularly to manage the complex mechanisms of turn-taking and turn-giving (Sacks *et al.* , 1978). There are logically speaking only four elementary functions related to turn management: (1) take turn (or floor grabber in the ICSI-MRDA terminology (Shriberg *et al.* ,

2004)); (2) maintain turn (or floor holder), a function often achieved by just speaking on; (3) give turn, often achieved by just stopping to speak; and (4) let the other speaker maintain her turn, often achieved by emitting short, approving noises called backchannels, such as *uh-huh* or *mhmm*.

The turn-managing function of utterances is related in some approaches to the grounding of speech acts, as for instance in Clark and Schaefer's (Clark & Schaefer, 1989) presentation/acceptance model mentioned in the previous section. Other approaches taxonomize the utterance-functions in this dimension based on analyzes of feedback (Allwood *et al.* , 1992), or based directly on the logical possibilities offered to speakers and hearers in interaction (Bunt, 2005, pages 6–7)[5].

While it is probable that every utterance plays some part in turn management, which is a dynamic phenomenon that is constantly under negotiation in a conversation, it is not clear whether some utterances have a default utterance-function in this dimension that could be left unmarked by a DA tagset. For instance, many utterances seem to function only as "continuers", when the current speaker maintains the turn by simply speaking on. Many other utterances, at the beginning of a turn, do not seem to use a specific device to take the turn other than simply starting to speak. These two cases are candidates for an unmarked utterance-function in the dimension of turn management. A marked utterance-function in this dimension could be assigned to the two other types (from the four described above), possibly specifying for each of them if it was successful or not. For instance, an attempt to take one's turn could succeed or not in interrupting the current speaker, but in both cases the function intended by the speaker is the same. In fact, it seems that the range of utterance-functions in this dimension is not so much a subject of debate in conversation analysis, compared to the precise manner of realizing them.

3.5 Adjacency pairs

Conversation analysts observed that utterances are often paired according to their functions, for instance a question and an answer (Schegloff & Sacks, 1973; Levinson, 1983, pages 303–308). Adjacency pairs have thus been defined as pairs of utterances that are adjacent, produced by different speakers, ordered as first part and second part, and typed—a particular type of first part requires a particular type of second part. First and second parts are more explicitly called, respectively, forward-looking and backward-looking function in the DAMSL tagset. Some of these constraints could be dropped to cover more cases of dependencies between utterances. For instance, re-

[5] An aspect related to turn-taking and turn-giving is the selection of the addressee, which can be ambiguous in multi-party meetings, and is partly coded, for instance, in the ICSI-MRDA DA tagset (Shriberg *et al.* , 2004). A proposal for full addressee coding in a corpus of multiparty dialogues was put forward by Jovanovic et al. (2005).

mote links between a first and a second part should be allowed, since other utterances can be sometimes inserted between them (e.g. a clarification subdialogue). Some utterances appear to play both roles, i.e. they are the second part in one adjacency pair (with respect to a previous utterance), and the first part in another one.

Adjacency pairs are relational by nature, but they could be reduced to labels ('first part', 'second part', 'none'), possibly augmented with a pointer towards the other member of the pair. Commonly observed types of adjacency pairs include the following ones, from Levinson (1983, page 336): request / offer / invite → accept / refuse; assess → agree / disagree; blame → denial / admission; question → answer; apology → downplay; thank → welcome; greeting → greeting.

One should not be confused by the terminological overlap between the domain of adjacency pairs and the domain of speech acts, which are distinct, albeit related dimensions of utterance-function. Adjacency pairs embody important functions of utterances that cannot be modelled through speech act theory, such as the notion of answer, which is not a speech act. According to Levinson (1983, page 293), "'answerhood' is a complex property composed of sequential location and topical coherence across two utterances, amongst other things; significantly, there is no proposed illocutionary force of answering." A similar observation of the fundamental difference between a question and an answer is made by Moeschler (2002, pages 241–243).

Several theories have set constraints on the acceptable second parts of adjacency pairs, thus leading to dialogue grammars, such as the models proposed by Sinclair and Coulthard (1975) or by the Geneva school (Roulet *et al.* , 1985; Moeschler, 1989). The observed variability of the relations between utterances seems to invalidate the possibility of normative sequencing rules proposed by such theories, and requires a more global approach to dialogue structure, as explained for instance by Moeschler (2002).

3.6 Topical organization of conversations

Studies in conversation analysis, starting with Schegloff and Sacks (1973), have concentrated on particular stages of conversations such as conventional openings and closings. Apart from these very specific stages, dialogues are often structured as series of thematic episodes, each of them characterized by one or more topics, as conceived by discourse analysis theorists (Brown & Yule, 1983, chapter 3). The following utterance-functions could be defined in this dimension: (1) open a conversation; (2) close a conversation; (3) start a new topic; (4) continue a topic; (5) end a topic. These functions seem to be mutually exclusive, though in some cases the same utterance could end one topic and begin the next one. The default function seems to be topic-continuer. More elaborate approaches to topical organization propose a hierarchy of topics and sub-topics, but the grounding of the notion of topic

or theme into more general theories of discourse and dialogue remains itself a matter of discussion (Wilson, 1998; Asher, 2004).

3.7 Rhetorical structure

The rhetorical roles that were defined in the Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) are mainly aimed at monologues, not dialogues. A study of discourse relations based on discourse semantics also proposed a similar typology of rhetorical relations, with a different grounding, which is more applicable to dialogues (Asher & Lascarides, 2003). Rhetorical roles, similarly to adjacency pairs, are a relational concept, concerning relations between utterances, not utterances in isolation. It is however possible, given that an utterance is a satellite with respect to a nucleus in only one relation, to assign to the utterance the label of the relation. This seems to require, in general, quite a deep analysis of dialogue structure.

The number of rhetorical relations defined in relation to the RST framework is quite variable, ranging from the 'dominates' and 'satisfaction-precedes' classes used by Grosz and Sidner (1986) to more than a hundred types, as summarized by Hovy and Maier (1995). The main relations defined by Mann and Thompson (1988) are: elaboration, circumstance, solutionhood, (non-)volitional cause / result, purpose, condition, interpretation, evaluation, restatement, summary, evidence, antithesis, concession, motivation, enablement, justification, background, contrast, joint, and sequence. A subset of these was used by Marcu (2000) in a series of experiments in automatic discourse parsing, aimed at text summarization. While such labels apply clearly to consecutive utterances in a turn, it is less clear how they apply to consecutive utterances from different speakers, or to utterances of the same speaker in different turns.

3.8 Politeness and other dimensions

Utterance-function in the sphere of politeness can be formalized using the notion of "face" or status proposed by Brown and Levinson (1987). Each utterance in a verbal interaction inevitably plays a role with respect to the negotiation of face, the default being probably here the neutral role, and other options being: (1) speaker attempts to save face (self-defence); (2) speaker attempts to save face of addressee (encouragement); (3) speaker threatens his own face (self-depreciation); and (4) speaker threatens face of addressee (aggression). Simpler categorizations of the politeness function could be: neutral vs. politeness-related, with the latter type possibly divided into positive (consensus) and negative (conflict).

Still other dimensions of utterance-function are sometimes reflected in DA tagsets used for computational applications. Some of them are not, in reality, related to utterance-function (e.g. intonation, or quotations of

previous utterances); others seem somewhat infrequent or difficult to detect automatically (e.g. humour or irony), or extra-linguistic (e.g. emotions) or non-functional (e.g. addressee coding).

3.9 Integrative and multi-dimensional theories

The analyses above strongly suggest that an utterance may serve several functions, in different planes. The integration of these planes has been the target of several theoretical approaches. For instance, Schiffrin (1987) considers the role of utterances in the exchange structure of a discourse, in its action structure, ideational structure, participation framework, and information state.

Bunt (2000) emphasizes three reasons why utterances can be multifunctional:[6] (1) indirectness, e.g. questions that function also as requests (see 3.3); (2) functional subsumption, e.g. promises are a specific case of informative statements; and (3) functional multidimensionality, since task-related functions are very often combined with dialogue control—e.g. feedback or turn management (Bunt, 2000, page 144). Bunt (2000) then proceeds to separate two types of functions: task-oriented vs. dialogue control. Each of these types is further sub-divided, down to the level of elementary functions such as 'agreement' or 'initiate-opening' [7].

The activity-based approach to pragmatics put forward by Allwood (2000) is in a certain sense an attempt to unify various utterance-functions, as introduced by his analysis of the following "background theories": Wittgenstein, speech acts, conversation analysis, Grice, dialogue grammars, Clark, and relevance theory. However, an integrated taxonomy applicable to DA tagging does not appear in the quoted study. The above-mentioned study of grounding by H. H. Clark is in fact an element of a more global theory of language use (Clark, 1996), which is more related to a plan-based approach to dialogue rather than a "shallower" approach based on DAs.

The multi-level scheme proposed by Traum and Hinkelman (1992) is probably the most computationally-oriented of the integrative theories of conversation acts proposed to date. First, the authors point out that different types of acts pertain to different granularity levels in dialogue. The following dimensions are considered: turn-taking (as studies in conversation analysis, at sub-utterance level), grounding (at utterance level), core speech acts (discourse units made of several utterances including an acknowledgment of the initial speech act presentation), and argumentation/rhetoric

[6] Apart of course from the case in which an utterance simply juxtaposes a series of atomic fragments, each carrying a specific function—this case is better modelled by considering that it is in fact the *turns* that are composed of *atomic* utterances.

[7] The two main subtypes of task-oriented communicative functions, originally proposed in (Bunt, 1989), are information-seeking and information-giving. The three main subtypes of dialogue control functions are feedback, interaction management and social obligations management, originally proposed in (Bunt, 1994).

structure (multiple discourse units). The application of this scheme to DA tagging seems complex, as it requires several levels of segmentation. However, turn-taking, grounding and speech acts could all be considered, by approximation, at the level of utterances: for turn-taking all that is needed is to consider the utterance that includes the sub-utterance fragment bearing the turn-taking function in the respective turn. For speech acts, one could consider only the initial attempt or presentation of the speech act (Traum & Hinkelman, 1992, section 2.1): if this is correctly grounded, then the grounding utterance itself would not accomplish any speech act. As we shall see, these three dimensions are reflected quite often in the four DA tagsets we consider below.

The Dynamic Interpretation Theory proposed by Bunt (2000; 2006) is embodied into a multi-dimensional tagset that was recently formalized in terms of eleven dimensions, such as task-related, auto-feedback, turn management, or social obligation management. These are accompanied by a set of general-purpose functions grouped for convenience into four classes, corresponding roughly to speech act types (information seeking, information providing, commissives and assertives). The communicative function of an utterance can either be described by a tag from one of the eleven particular dimensions, or by one of the general-purpose tags accompanied by a semantic content in one of the eleven dimensions. The DIT++ tagset contains thus eleven particular dimensions, and, formally, each of them contains dimension-specific tags and a copy of the set of general-purpose tags.

In a recent synthesis oriented towards language processing, Jurafsky and Martin mention the influence on dialogue modelling of the following approaches: conversation analysis, speech acts, grounding, conversation structure and implicatures (2000, chapter 19). A survey conducted by Samuel (1999, Appendix G) displays a much wider range of theoretical and applicative proposals for taxonomizing utterance-function.

3.10 Global approaches to communication

A number of theories of language-based communication could also be applied to the problem of categorizing utterance-function. Relevance Theory (Sperber & Wilson, 1986/95) proposes an ostensive-inferential model of communication which replaces the pre- and post-conditions on speech acts with the inferential effects of an utterance on the receiver's set of beliefs. The theory of Generalized Conversational Implicatures (Levinson, 2000) proposes an alternative solution to the problem of indirect speech acts. The Segmented Discourse Representation Theory (Asher & Lascarides, 2003) was exploited by its authors to integrate some of the dimensions cited above, namely speech acts, rhetorical relations, adjacency pairs, and thematic annotation (Asher & Lascarides, 2001; Asher, 2004), and to provide a framework for relations between DAs. The dimensions discussed above (3.3–3.8) could thus appear

one day as particular projections of a unique global theory. However, the candidate theories are still too complex or too abstract to allow such developments, though some of them seem computationally tractable (Schlangen *et al.* , 2001).

4 A Series of DA Tagsets Used for Automatic Tagging of Large Corpora

We now turn to four DA tagsets that illustrate the difficulty of choosing between a one-dimensional and a multi-dimensional tagset. Compared to the many other DA tagsets that have been proposed (Klein & Soria, 1998; Samuel, 1999, Appendix G), these tagsets are among the few general-domain tagsets (i.e., not dedicated to dialogues for a specific task or domain) that have been used to annotate large scale resources (see Klein & Soria (1998) for a synthesis), as opposed for instance to the tagsets used for the Map-Task Carletta *et al.* (1997) or the Verbmobil Jekat *et al.* (1995) corpora. Therefore, frequency distributions, inter-annotator agreement scores and automatic tagging scores are more readily accessible and more reliable than for other tagsets.

In addition, the historical relatedness of the four tagsets offers insights on the respective roles of theoretical inspiration and practical considerations from one tagset to another. The DAMSL tagset initially exploited dialogue theories in several dimensions, while the SWBD-DAMSL tagset was derived from DAMSL as a one-dimensional tagset based on frequency information of occurring DAMSL labels in a corpus of telephone conversations. Later on, the ICSI-MRDA tagset made use of combinations of SWBD-DAMSL tags into one label, and MALTUS attempted to reduce again the number of ICSI-MRDA tags by grouping tags into classes and making explicit some mutual-exclusiveness constraints observed on a corpus of multi-party meetings. Both DAMSL and MALTUS exploit the observation that only a very small proportion of the combinations of tags theoretically allowed by multi-dimensional tags do actually occur in reality.

4.1 DAMSL

The Dialogue Act Markup in Several Layers (DAMSL) was proposed by the Discourse Resource Initiative (Allen & Core, 1997; Core & Allen, 1997). Though initially proposed as a general dialogue annotation scheme, DAMSL focuses to a certain extent on task-oriented dialogues[8].

[8] According to Jurafsky and Martin (2000, page 729), DAMSL is "focused somewhat on the kind of dialogue acts that tend to occur in task-oriented dialogue." In their revised version of the chapter, Jurafsky and Martin maintain that "DAMSL is focused on task-oriented dialogue."

DAMSL draws inspiration from theoretical dimensions such as speech acts (including their grounding) and adjacency pairs or feedback, while attempting to synthesize existing schemes as well[9]. DAMSL distinguishes four categories of utterance-functions: communicative status, information level, forward-looking function and backward-looking function. These are not 'dimensions' in the above sense because each utterance may be tagged with as many DAMSL functions as needed, from as many dimensions as needed. We estimated elsewhere (Popescu-Belis, 2003) that there are about 4 million possible combinations of DAMSL tags into 'labels' (as defined earlier), even with the hypothesis that most tags in the lowest-level sub-categories are in fact mutually exclusive, as one can infer from their meaning, for instance 'accept' vs. 'maybe' vs. 'reject' in the backward-looking/agreement sub-category.

The generality of the DAMSL scheme makes it adaptable to many types of tasks and dialogues, e.g. collaborative human-human dialogues (Di Eugenio *et al.* , 1998, 2000) or information-seeking human-computer dialogues (Walker & Passonneau, 2001; Rosset & Lamel, 2004). To our knowledge, there is much less available data annotated with the derived tagsets than with SWBD-DAMSL itself.

4.2 SWBD-DAMSL

The application of DAMSL to the two-party telephone conversations of the Switchboard corpus has led to the development of the one-dimensional SWBD-DAMSL tagset (Jurafsky *et al.* , 1997), aimed in particular at automatic DA tagging[10]. About 200,000 utterances were first manually annotated using DAMSL. It was then observed that only 220 different combinations of tags occurred in the data (Jurafsky *et al.* , 1998). These 220 labels were further clustered into 42 synthetic tags such as statement (36% of the utterances), opinion (13%), agree/accept (5%), yes-no-question (2%). Therefore, although SWBD-DAMSL has slightly more individual tags than DAMSL, the fact that no combinations of tags are allowed for an utterance results in a much smaller search space for automatic DA tagging: 42 tags compared to the 4 million possible combinations of DAMSL tags. This considerable reduction of the search space is typical of one-dimensional tagsets.

[9] According to Jurafsky and Martin (2000, page 729), "DAMSL codes [...] various levels of dialogue information about utterances. Two of these levels, the forward looking function and the backward looking function, are extensions of speech acts which draw on notions of dialogue structure like the adjacency pairs [...] as well as notions of grounding and repair."

[10] The SWBD-DAMSL tags were also used as features of language models for speech recognition on the Switchboard data (Stolcke *et al.* , 2000).

4.3 ICSI-MRDA

The annotation of utterance-functions on the ICSI Meeting Recorder corpus (Janin *et al.* , 2003; Morgan *et al.* , 2003) required some modifications of the SWBD-DAMSL tagset for an application to multi-party dialogues. The ICSI-MRDA tagset therefore extended the SWBD-DAMSL tagset, and at the same time removed the mutual-exclusiveness constraint: each utterance bears a label made of as many SWBD-DAMSL tags as applicable (Dhillon *et al.* , 2004; Shriberg *et al.* , 2004). So, while SWBD-DAMSL was an attempt to reduce the dimensionality of DAMSL, ICSI-MRDA allows the combination of SWBD-DAMSL tags, instead of reusing the original DAMSL tags[11].

ICSI-MRDA allows the annotation of many fine-grained distinctions, for instance between agreement (or acceptance), acknowledgment, backchannel, and floor grabber, as expressed by various discourse particles such as *yeah*, *right*, *okay*, and *uh-huh* (Bhagat *et al.* , 2003). The tagset also allows the annotation of disruptions (interrupted or abandoned utterances) or turn-taking mechanisms (floor grabber or floor holder). To increase inter-annotator agreement (see Section 7.2 below), a number of abstractions of the tagset called *class maps* were defined, grouping tags into higher-level categories. For instance, the broadest class map reduces all ICSI-MRDA to only five high-level tags: statement, question, backchannel, floor-holder, and disruption.

4.4 MALTUS

We proposed the MALTUS tagset (Multidimensional Abstract Layered Tagset for Utterances) in order to reduce the number of possible ICSI-MRDA labels, by grouping some tags into classes and assigning mutual-exclusiveness constraints between them (Popescu-Belis, 2003; Clark & Popescu-Belis, 2004). MALTUS is thus an abstraction of ICSI-MRDA, akin to a constrained class map, and therefore remains compatible with the ICSI-MRDA tagset and allows the reuse of ICSI-MRDA annotated meetings, on which it is based.

The following tags are used in MALTUS: an utterance is either undecipherable (U) or it has exactly one general tag and zero or more specific tags; it can also bear a disruption mark. There are four general tags: statement (S), question (Q), backchannel (B), floor-holder/grabber (H). Specific tags are: positive/negative/other response (RP/RN/RU); attention-related (AT), such as an understanding check or an acknowledgment; performative (DO), such as a suggestion or a commitment; politeness-related (PO); and restated information (RI).

[11] The authors seem to have overlooked the fact that the considerable increase of the search space may render the ICSI-MRDA tagset unsuitable for automatic tagging, which seems nevertheless to be the underlying goal of their study (Dhillon *et al.* , 2004, page 4). The freedom to use as many SWBD-DAMSL as necessary for characterizing an utterance makes ICSI-MRDA more suitable for the empirical study of multi-party meetings.

MALTUS maps straightforwardly the ICSI-MRDA classes of tags, and correspondences to other tagsets were also defined (Popescu-Belis, 2003), from the more specific tagsets (ICSI-MRDA or SWBD-DAMSL or DAMSL) to the more abstract one (MALTUS). When used for converting the annotation of an existing resource, these correspondences must be used along with the mutual-exclusiveness constraints between tags.

5 What is Dimensionality?

5.1 Definition

In this section, we provide a mathematical formulation of the differences between one-dimensional and multi-dimensional tagsets. Our model accommodates a variety of guidelines that appear in existing tagsets, in particular multi-dimensional ones. A more detailed model is proposed by Bunt (2005), in an attempt to normalize the definition of DA tagsets[12].

From a theoretical perspective, a one-dimensional tagset is a set $A = \{a_1, a_2, \ldots, a_N\}$, each utterance being tagged with exactly one elementary tag $a_n \in T$. A multi-dimensional tagset is a collection of dimensions (or classes, categories, etc.) $\mathcal{T} = \{A, B, \ldots\}$ where each dimension is in turn a list of tags, say $A = \{a_1, a_2, \ldots, a_M\}$, $B = \{b_1, b_2, \ldots, b_N\}$, etc. When a multi-dimensional tagset is used, each utterance is tagged with a composite *label* or tuple of tags $(a_i, b_j, \ldots)$. In this idealized view, each utterance receives exactly one tag from each dimension, a fact that is often indicated by the creators of the tagset as "pick *exactly* one tag from dimension A (or class/category), exactly one tag from dimension B, etc." If the multi-dimensional tagset is accompanied by the instruction "pick *at most* one tag from dimension A", then the notations we introduced remain valid provided the empty tag $\emptyset$ is added to the concerned dimension[13].

Tagsets accompanied by the instruction "pick *all* tags that apply from dimension A" are also compatible with our definition. From our perspective, in this case, dimension A is no longer a single theoretical dimension, but is modelled by a set of binary dimensions: $\{a_1, \emptyset\} \times \{a_2, \emptyset\} \times \ldots$ since this set product generates all possible combinations of tags from A (without repetition). If the instruction states "pick at least one tag" from dimension A, then the empty label $(\emptyset, \emptyset, \ldots, \emptyset)$ should be excluded from the set product $\{a_1, \emptyset\} \times \{a_2, \emptyset\} \times \ldots$ corresponding to A.

[12] Bunt's model includes also the notion of categories of DAs, which is useful for clarifying the description of a tagset, but does not play a role in the theoretical arguments we discuss here.

[13] Note that if all dimensions are accompanied by the instruction "pick at most one tag", then the option that no tag is selected in any dimension (empty label) is theoretically possible, but should be ruled out. Hence the exact set of all possible labels should be the set product minus the empty tuple: $(A \times B \times C \times \ldots) \setminus \{(\emptyset, \emptyset, \emptyset, \ldots)\}$.

A consequence of the difference between one- and multi-dimensional tagsets as defined here is, in practice, the magnitude of the number of possible labels. One-dimensional tagsets usually list all the possible labels as an unstructured set, therefore they rarely have more than a hundred tags (e.g. SWBD-DAMSL, see 4.2 below). However, if a comparable number of tags appears in the dimensions of a multi-dimensional tagset, then the number of possible labels is the product of the sizes of the dimensions, usually a much larger number. If a multi-dimensional tagset has N dimensions, each of size k_i ($1 \leq i \leq N$), then the size of the tagset is $k_1 \times k_2 \times \ldots \times k_N$ (or k^N if each dimensions has k tags), which is potentially very large. However, if these tags were organized as a one-dimensional tagset, that is if the tags were mutually exclusive, then there would be only $k_1 + k_2 + \ldots + k_N$ (or $N \times k$) possible labels (i.e. tags).

5.2 Linearization of multi-dimensional tagsets

A multi-dimensional tagset $\mathcal{T}$ can be formally *linearized* by listing, in an unstructured set, all the possible labels or complex tags $\{\ldots, (a_i, b_j, \ldots), \ldots\}$. The size of the resulting one-dimensional tagset, i.e. the total number of mutually-exclusive labels, is the product of the sizes of each dimension of the original tagset. This formal equivalence shows that dimensions are mainly relevant to the conceptual organization or description of a tagset, and that they greatly reduce the number of labels that must be enumerated.

More concretely, consider the following example of a very simple two-dimensional tagset: $\mathcal{T} = \{A, B\}$, where $A = \{a_1, a_2\}$, $T_2 = \{b_1, b_2\}$, and a_1 and a_2 are mutually-exclusive tags, as are b_1 and b_2. The possible labels are the (a_i, b_j) couples, therefore $\mathcal{T}$ is formally equivalent to $\mathcal{T}' = \{(a_1, b_1), (a_1, b_2), (a_2, b_1), (a_2, b_2)\}$. In this case, the linearized version is as simple to enumerate as the multi-dimensional one, but this is generally not the case: if A has α tags and B has β tags, then $\mathcal{T}$ can be enumerated using $\alpha + \beta$ tags, while $\mathcal{T}'$ contains $\alpha \times \beta$ tags.

5.3 Independence of dimensions

Dimensions can be independent or not: absolutely—if any combination of tags is possible, a fact that can sometimes be verified empirically—or statistically, if the probability of any combination of tags is the product of the probabilities of each tag.

The use of dimensions in the description of a tagset facilitates the expression of constraints on the combinations of elementary tags, simply by cancelling some combinations of tags from the list of complex labels. A more radical option is to group incompatible tags into one dimension, with the instruction "pick exactly one"—tags can be incompatible for instance because they reflect alternative utterance-functions in a given theory.

For instance, in the above example, if one knew that a_2 and b_1 were incompatible tags, then the (a_2, b_1) combination could be removed from the one-dimensional form of the tagset, or ruled out by an explicit constraint in the two-dimensional version. If the A and B dimensions contained ten tags each, then it would be much easier to define the tagset using the two dimensions rather than enumerating the 100 labels of the one-dimension $A \times B$ tagset.

Dimensions can be independent or not: according to Bunt (Bunt, 2005, page 7), "two dimensions [...] are independent if any pair of tags from the two dimensions is admissible." It is not often the case, however, that impossible combinations of tags are stated from the beginning in the annotation guidelines for multi-dimensional tagsets—because such constraints could be very numerous (for combinatorial reasons) and they are not known a priori. More often (see 4.2 below) such constraints are derived from empirical observations on large-scale annotated corpora. In addition, the stronger notion of *statistical* independence between dimensions could be proposed (when the probability of any combination of tags is the product of the individual probabilities of each tag). It seems however uneasy to find two fully independent dimensions of utterance-function in the statistical sense.

6 Dimensionality of Tagsets used for Automatic Tagging

We now consider the factors that influence the performance of taggers which use machine learning; we then derive from these factors some arguments regarding the dimensionality of DA tagsets.

The performance of automatic DA taggers must be interpreted with respect to human performance on identifying DAs. Humans do not have 100% accuracy, partly because of performance errors, and partly because some utterances are intrinsically ambiguous. According to some analyses quoted above (Section 3.3), the DA carried by an utterance is a form of joint action between the speaker and the hearer, that is, the DA is fully realized only after an initial utterance has been grounded by the receiver through specific signals (Clark & Schaefer, 1989; Traum & Hinkelman, 1992). The accuracy level required from an automatic DA classifier could either be compared to the accuracy of a human interlocutor in a dialogue (online "DA prediction" task), or to inter-annotator agreement as discussed in Section 7.2 (offline "DA annotation" task—in both cases humans are below 100% accuracy.

6.1 Limiting factors for statistically-based DA taggers

DA taggers are classifiers that sort utterances from a given set into a number of classes corresponding to the DA tagset. From each utterance a number

of features are extracted (generally not by the tagger itself but by a pre-processor), and then used for classification. The DA tagger can use machine learning to train its classification model, i.e. learn a statistical correlation between observed features and corresponding classes.

In this view, three principal factors that influence the performance of automatic DA taggers can be identified: (1) the amount of data available; (2) the size of search space, i.e. the number of classes (tags/labels) to choose from when tagging an utterance (this factor is not specific to statistically-based methods); (3) the number and nature of the features used for the classification of utterances.

The general view is that more data and fewer classes both increase classification accuracy, that using too many features decreases accuracy (if data set remains below the size needed to reach asymptotic performance) and that omission of relevant features decreases accuracy as well. These three factors are in fact related: for instance, the desirable amount of data depends on the number of features and on the number of classes.

In addition, the evaluation measures can influence the performance of a DA tagger and the choice of a tagset. Here, we assume that evaluation consists simply in checking whether for each utterance, the DA label assigned by a system is identical to the one assigned by humans, but more complex scoring schemes can give a different view of performance (Lesch *et al.* , 2005a). For instance, when labels are composed of tags, the amount of overlap between the system's label and the correct one can be used to compute a non-binary score.

Several arguments based on the above considerations can be constructed either in favour or against multi-dimensional DA tagsets. To illustrate the arguments, we will consider a two-dimensional tagset $\mathcal{T} = \{A, B\}$ with $A = \{a_1, a_2, \ldots, a_m\}$ and $B = \{b_1, b_2, \ldots, b_n\}$. This is formally equivalent to the one-dimensional joint tagset $A \times B = \{(a_1, b_1), \ldots\}$ which has $m \times n$ mutually-exclusive tags.

6.2 Role of the available training data: asymptotic performance

Humans make use of many sources of information and of complex linguistic and pragmatic knowledge in order to assign a DA to an utterance. It is likely that for a significant number of utterances, DA assignment cannot be done without such information/knowledge, though this question has rarely been studied empirically[14]. Therefore, we may state that a computer program that is capable to use semantic representations of utterances, contextual information and complex world knowledge would have no more trouble than

[14]One could hypothesize that since the actual DA tagging methods, which do not make use of such complex features, reach about 70-80% accuracy, there are about 20% of the utterances which require complex information/knowledge for correct tagging.

a human identifying the correct DA carried by an utterance[15]. However, encoding the full semantic complexity of utterances as features available to machine learning systems is not a current possibility. Many DA taggers use in fact only a limited set of features, which means that these features are insufficient to disambiguate the DA of all utterances[16]. Even if unlimited hand-labelled data was available for training, such a classifier would only reach a score well below 100%, since the features it used simply do not allow for a full disambiguation of DAs—this score is called *asymptotic performance*.

How far is the available training data from what is needed for asymptotic performance? The required amount of training data depends, among other things, on its number of degrees of freedom. For DAs, this is related to the number of possible words and n-grams used for DA tagging, hence at least thousands, plus thousands of thousands ... According to Ng and Jordan (2001), "the number of examples needed to fit a model is often roughly linear in the number of free parameters of a model. This has its theoretical basis in the observation that for 'many' models, the VC dimension is roughly linear or at most some low-order polynomial in the number of parameters [...] and it is known that sample complexity *in the discriminative setting* is linear in the VC dimension (Vapnik, 1998)." The authors show that another class of models, called *generative*, among which the Naive Bayes classifier, "approach [their] asymptotic error much faster than the discriminative model—possibly with a number of training examples that is only logarithmic, rather than linear, in the number of parameters."

Therefore, the size of the training set should be proportional to the number of degrees of freedom (or to its logarithm if generative models are used) in order to attain asymptotic performance of the DA tagger. However, it seems that the number of degrees of freedom, if words, bi-grams and tri-grams are used, is anyway much greater that the size of available DA-tagged corpora, hence asymptotic performance is far from being reached at present[17].

6.3 Role of the size of search space

One-dimensional tagsets, because they are defined extensionally, by the enumeration of all possible tags, tend to limit the number of tags to less than one hundred. Conversely, multi-dimensional tags, if insufficient constraints are expressed on the combinations of dimensions, can generate thousands

[15]Provided of course that the knowledge is "optimally" used, which is the case for a system based on machine learning which has enough training data to learn the impact on all features for classification.

[16]The features are, quite often, words as indicators, n-grams ($n < 3$), surrounding n-grams, and preceding labels.

[17]If only very few features (such as cue words) were used, asymptotic performance could be reached but would be quite low anyway.

or even millions of combinations, as shown above. So, multi-dimensional tagsets tend to provide very huge search spaces if they are not constrained either by using mutually exclusive-dimensions or constraints on tag combinations. Such constraints could be learned empirically by the system, but the required hand-labelled data to observe all of them is far greater than resources available today. Therefore, if non-empirical observations or a priori analyses show that some combinations of tags cannot occur, then such constraints could be used to reduce considerably the search space.

The smaller size of the search space is probably the main factor explaining why joint classifiers, which consider only the combinations of tags observed in the training data, outperform the combination of dimension-specific classifiers, which behave as if all possible combinations of tags could indeed occur—as shown in Section 7.3 below. However, Manning and Schütze (1999, p.144–145) discuss the possibility that a larger tagset (for POS tagging) actually increases tagging accuracy, if it has greater "predictive" power with respect to the sequence of tags[18].

6.4 Role of the features used for classification

The (non)independence of the features extracted from the training data is another factor that influences tagging accuracy. However, the analysis at this level depends strongly on the classification algorithm, the hypothesized relevance of the features that are used, and the training data available. For instance, some results show that Naive Bayes classifiers, which assume independence of the features, still have good performance even when this is not true (Domingos & Pazzani, 1996, 1997)[19]. So, even though one would expect the accuracy to decrease when features are not independent, this may

[18] In their considerations on the design of a POS tagset, Manning and Schütze argue that a smaller search space does not always increase accuracy: "a tagset encodes both the target feature of classification, telling the user the useful information about the grammatical class of a word, and the predictive features, encoding features that will be useful in predicting the behaviour of other words in the context. These two tasks should overlap, but they are not necessarily identical.[...] So long as the same tagset is used for prediction and classification, making such changes tends to be a two-edged sword: splitting tags to capture useful distinctions gives improved information for prediction, but makes the classification task harder. (Note: This is unless one category merges two very separate distributional clusters, in which case splitting the category can actually sometimes make classification easier.) For this reason, there is not necessarily a simple relationship between tag set size and the performance of automatic taggers" (1999, p.144–145).

[19] The probabilities of the classes are in this case quite wrong, but the comparison between them, on which the classification is based, is preserved: "...the [Naive Based classifier] does not in fact assume attribute independence, and can be optimal even when this assumption is violated by a wide margin. The key to this finding lies in the distinction between classification and probability estimation: correct classification can be achieved even when the probability estimates used contain large errors. We show that the previously-assumed region of optimality of the [Naive Based Classifier] is a second-order infinitesimal fraction of the actual one." (Domingos & Pazzani, 1996).

not happen for Naive Bayes classifiers. In fact, Maximum Entropy classifiers seem to be used more frequently for DA tagging (Ang *et al.* , 2005; Lesch *et al.* , 2005b), but the analyses presented by Klein and Manning (2003) tend to show that they are not sensitive to correlated features, when enough training data is available (in addition, they make it easy to incorporate information about the correlations between classes).

6.5 Dimensionality of tagsets in relation to dimensionality of classifiers

Given a multi-dimensional DA tagset (say $\mathcal{T} = \{A, B\}$), is it preferable to construct classifiers for each dimension, or to use a joint classifier for $A \times B$? This determines the choice between a one-dimensional and a multi-dimensional tagset, since separate classifiers can only be built for multi-dimensional tagsets. For a one-dimensional tagset, only one classifier can be constructed, unless the tagset is partitioned artificially into several dimensions. Formally, there two possible ways to construct DA taggers for the two-dimensional tagset $\mathcal{T}$: (1) train two separate classifiers, $\mathcal{C}_A$ for A and $\mathcal{C}_B$ for B, then combine them into a resulting classifier $\mathcal{C}_A\mathcal{C}_B$; or, (2) train a joint classifier $\mathcal{C}_{AB}$ on the product set $A \times B$.

An argument in favour of the joint classifier $\mathcal{C}_{AB}$ is that its search space tends to be much smaller than the search space of independent ones if certain combinations of tags are never seen in the training data. In other words, the dependencies between tags (such as two tags that never co-occur) are internalized in the joint classifier, and much more difficult to incorporate in the combined independent classifiers $\mathcal{C}_A\mathcal{C}_B$. If A and B are not independent, one would expect $\mathcal{C}_{AB}$ to perform better than $\mathcal{C}_A\mathcal{C}_B$ since $\mathcal{C}_{AB}$ is aware of the combinations of tags that occur more (or less) frequently than predicted by observations that are restricted to A and B separately[20].

However, an opposite argument in favour of independent classifiers can be constructed on a different basis. At present, given the available amounts of training data, the taggers are still far from reaching asymptotic performance. Their accuracy could therefore increase if more training data was available, or alternatively, if the number of degrees of freedom was reduced. Or, put differently, asymptotic performance is reached with less data when the number of degrees of freedom is smaller. Indeed, if independent classifiers ($\mathcal{C}_A\mathcal{C}_B$) are used, the feature sets for each classifier are smaller than those for the joint one, which necessarily uses all the features. Moreover, if the various dimensions of a multi-dimensional DA tagset represent different conversation phenomena, it is very likely that the corresponding classifiers

[20] If A and B are statistically independent dimensions, then the *asymptotic* performance of the two classifiers $\mathcal{C}_A\mathcal{C}_B$ and $\mathcal{C}_{AB}$ should be the same. The exact proof depends on the classification models that are used—see the Appendix for an analysis using an idealized model.

use quite different features or indicators. These features are at least partially determined by a priori knowledge of the phenomena, even if they are of course refined during training[21].

Therefore, one would expect the independent classifiers to reach a higher combined score, since in proportion to the feature set, each classifier has more training data available—assuming that the size of the hand-labelled corpus cannot be increased at will. Empirical results from experiments with automatic taggers, outlined in Section 7.3 below, are thus necessary to assess the relative importance of dimensionality choices, for DA tagsets and related classification methods.

7 Empirical Data on Large Scale Resources

When large dialogue corpora are annotated with DAs, and used as training and test data for automatic tagging, a number of empirical results help answering our main question: are there any reasons to prefer either multi-dimensional tagsets or one-dimensional ones? In this section, we examine DAs frequencies, inter-annotator agreement scores and automatic tagging scores, with regard to the dimensionality problem.

7.1 Theoretical vs. observed DA labels

We have pointed out earlier in Section 4.2 that Jurafsky et al. (1998) found that only 220 different combinations of DAMSL tags occurred in the 200,000 Switchboard utterances, while we estimated the theoretical number of possible combinations at about 4 million (Popescu-Belis, 2003). Similarly, we have shown that the number of possible ICSI-MRDA composite labels reaches several millions, even without considering disruption marks and allowing a maximum of six tags per label (as observed empirically)—see the upper part of Table 7.1. But we also observed that in the 75 ICSI-MRDA meetings, segmented in 113,560 atomic utterances, only 776 different labels occurred, mostly composed of 1, 2 or 3 tags (ca. 70%). In other words, the 113,560 tokens of composite labels corresponded to only 776 types.

The last column of Table 7.1 using an ideal (oracle) tagger that is limited to DA labels with a maximum of N tags. Depending on the number of tags per label (N), the numbers of theoretical vs. observed label types increase,

[21] For instance, suppose that two mutually-exclusive tags are Q (question) and B (backchannel), and that the four following features are used: pitch variation along the utterance (f_1), presence of an interrogative pronoun (f_2), duration of utterance (f_3) and detected activity of other speakers (f_4). The joint classifier for {Q, B}, noted $\mathcal{C}_{\mathrm{QB}}$, uses $\{f_1, f_2, f_3, f_4\}$, while separate classifiers for questions {Q, ¬Q} and backchannels {B, ¬B}, noted $\mathcal{C}_{\mathrm{Q}}$ and $\mathcal{C}_{\mathrm{B}}$, could use only $\{f_1, f_2\}$ and respectively $\{f_3, f_4\}$. The number of degrees of freedom is divided by two, and so is the amount of data needed to reach asymptotic performance.

Tagset	Tags/ label (N)	Theoretical (possible) label types	Actual (observed) label types	Actual (observed) label tokens	Max.acc. with N tags
ICSI-MRDA	1	11	11	68,213	0.6007
	2	429	129	37,889	0.9343
	3	8,151	402	5,054	0.9788
	4	100,529	176	2,064	0.9970
	5	904,761	49	326	0.9999
	6	6,333,327	9	14	1.0000
Total	–	7,347,208	776	113,560	–
MALTUS	1	4	4	84,092	0.74051
	2	28	14	28,366	0.99003
	3	72	29	1,089	0.99997
	4	88	3	3	1.00000
Total	–	192	50	113,560	–

Table 1: Number of theoretical and observed ICSI-MRDA and MALTUS labels on the ICSI-MR corpus with 113,560 atomic utterances.

then decrease, at very different rates. There are no observed labels bearing more than six ICSI-MRDA tags, or more than four MALTUS tags. Hence, a tagger using at most six ICSI-MRDA tags can in principle reach 100% accuracy, but a tagger using at most four ICSI-MRDA tags is necessarily limited at 99.70% on the current ICSI-MR corpus. Or, for instance, if only ICSI-MRDA labels made of 1, 2 or 3 tags were used for automatic tagging, only about 2% of the utterances would be intrinsically impossible to tag correctly, but the search space would be reduced from several million combinations to less than 10,000.

Therefore, a reasonable solution in order to limit the search space for automatic tagging is to consider only labels with a small number of tags. In the case of ICSI-MRDA, figures in Table 7.1 show that when considering search space made of only 1, 2, and 3-tag labels, of which less than 600 types do occur, the lower bound on the recognition error would be at most 2%, *if* the test data was the actual ICSI-MRDA data (i.e. less than 2% of the utterances bear four tags or more).

Similarly to SWBD-DAMSL, the size of MALTUS is several orders of magnitude smaller than the number of possible combinations of DAMSL or ICSI-MRDA tags. There are indeed no more than 600 possible MALTUS labels (see upper part of Table 7.1), or 200 without disruptions. However, only 50 MALTUS labels appear in the 113,560 utterances of the converted ICSI-MRDA data, and only 22 labels occur more than 20 times each (see Table 2). In fact, only 136 utterances (0.12% of the total) bear other labels than the most frequent 22 ones. In case only these 22 MALTUS labels were used for

Label	Nb. occ.	Label	Nb. occ.	Label	Nb. occ.
B	15,180	S	51,304	S^RP	7,612
H	12,288	S^AT	8,280	S^RP^DO	41
Q	5,320	S^AT^RI	273	S^RP^RI	436
Q^AT	3,137	S^DO	3,935	S^RN	2,219
Q^AT^RI	69	S^DO^RI	32	S^RN^DO	38
Q^DO	239	S^PO	791	S^RN^RI	46
Q^RI	60	S^RI	765	S^RU	1,298
				S^RU^PO	61

Table 2: MALTUS labels occurring more than twenty times.

automatic tagging of the ICSI-MRDA data, to reduce search space, the maximal theoretical accuracy would be 99.88%. This not only far above the actual performances of DA taggers, but also well above the observed values of inter-annotator agreement (see Section 7.2 below)).

It appears from Table 2 that the vast majority of the utterances bear only one elementary function, such as statement, backchannel, floor-holder/grabber, or question. Next by order of frequency are attention-related labels such as 'Q^AT' (e.g. tag questions) or 'S^AT' (e.g. acknowledgments). The next types, by order of frequency, are positive, negative, and undecided responses, which are in some cases followed by modifiers (restated information, politeness, performative). Statements accompanied by performatives (e.g. commitments, suggestions) are in the same frequency range.

The frequency of each DA type per meeting is a characteristic of the ICSI-MR corpus and of its dialogue genre. For instance, about 45% of the utterances of a meeting are on average pure statements (bearing only the 'S' tag), the 95% confidence interval being only ±1%. At the opposite side of the frequency spectrum, an infrequent label such as 'S^DO^RI' is applied to only 0.03% of the utterances, with a much larger confidence interval, ±0.01% (absolute value). Similarly, 'Q^DO ' has an average frequency of 0.21%, with a confidence interval of ±4% (absolute value)[22].

7.2 Effects on manual tagging

Inter-annotator agreement scores decrease with the size of the tagset, but precise experiments to assess this fact are quite costly (Carletta *et al.* , 1997). A widespread metric of inter-annotator agreement is Cohen's κ (*kappa*), a similarity score that factors out agreement by chance; agreement is often considered acceptable, following Krippendorff's account, only when $\kappa > 0.8$ (Cohen, 1960; Krippendorff, 1980; Carletta, 1996)—see also recent discussions about the validity of these metrics (Di Eugenio & Glass, 2004; Craggs

[22]Confidence intervals are computed here using the observed standard deviation and Student's law, but the use of the binomial law provides similar figures.

& McGee Wood, 2005).

Most cases of acceptable inter-annotator agreement scores occur for low-dimension tagsets: for instance, $\kappa = 0.8$ for SWBD-DAMSL (Jurafsky *et al.* , 1997). To reach comparable reliability ($\kappa = 0.79$) on the ICSI-MRDA tagset, the authors had to apply a class map reducing the tagset to only five abstract labels[23]; using a more detailed one (with about 15 tags), κ decreases to 0.76 (Shriberg *et al.* , 2004, p.99). Inter-annotator agreement for both MapTask and VerbMobil tagsets is 0.83 (Klein & Soria, 1998).

The results put forward by Doran at al. (2003, page 136) explicitly show that inter-annotator agreement decreases when the complexity of the tagset increases. In their experiments with a one-dimensional tagset, subset of the CSTAR tagset (CSTAR, 1998), κ reached 0.90 for a 20-tag subset, and only 0.71 for a 26-tag subset.

In another empirical study, Di Eugenio et al. (1998; 2000) studied inter-annotator agreement on about 500 utterances, with a tagset similar to DAMSL, providing κ values for different dimensions. These values vary from 0.83 for the "Assert / Reassert / NIL" dimension and 0.79 for "Answer / NIL", to 0.72 for both "Open-Option / Info-Request / Action-Directive / NIL" and "Offer / Commit / NIL", to 0.54 (the lowest figure) for "Accept / Reject / Hold / NIL". Reliability thus decreases when the number of possible tags increases, and some dimensions appear to be more difficult to categorize than others. These figures are slightly better than those obtained with DAMSL, which were, respectively: 0.68, 0.81, 0.71, N/S (not significant) and 0.81 (Core & Allen, 1997, Tables 2 and 3), probably due to a better adaptation of the annotation guidelines to the type of data. Again, annotator agreement generally decreases as the number of possible tags increases, but some dimensions appear to be more difficult to categorize than others regardless of their number of tags. Note also that the scores would be even lower if reliability was scored jointly on the five dimensions, i.e. by testing whether the complex labels are identical across annotators: indeed, for DAMSL, the overall κ is only 0.56.

A recent experiment with the DIT++ tagset has also investigated dimension-specific inter-annotator agreement Geertzen & Bunt (2006). Looking only at dimensions for which more than 100 pairs of annotations were done, it appears that the highest value of *kappa*, $\kappa = 0.82$, was reached for the turn management dimension with only four tags (general-purpose tags are apparently not used in this dimension). However, only 34% of the pairs containing at least one auto-feedback tag are in fact annotated with such a tag by *both* annotators, a fact that is not considered for computing κ in this study. For the task-related dimension, for which more than 40 general-purpose tags are used, $\kappa = 0.47$, a much lower value, though it can be corrected to 0.71 if the distance between tags from the same branches of the hierarchy is taken into

[23] Statement, question, floor (turn) management, backchannel, disruption.

account in the score, as Geertzen and Bunt propose. The agreement for the auto-feedback dimension is even lower, at $\kappa = 0.21$ (corrected at 0.57), while the number of tags for this dimension appears to be higher (general-purpose ones and ten dimension-specific ones).

However, unlike the previous results, all these agreement figures are dimension-specific, hence quite higher than the overall DAMSL or DIT++ reliability that was scored by testing whether the complex labels of each utterance are identical across annotators. As dimension-specific errors tend to cumulate, overall reliability is lower—for instance this is estimated at $\kappa = 0.56$ using the DAMSL on an unspecified data set Klein & Soria (1998). Therefore, depending on what level of annotation reliability is sought, the use of complex, multi-dimensional DA tagsets seems to require supplementary inter-annotator verification.

7.3 Effects on automatic tagging

The factors that influence tagging accuracy, discussed in Sections 6.2–6.4 above, are further constrained by the specific context of each study. For instance, the amount and reliability of available hand-labelled data is often limited by the cost of such a resource—although a more efficient, selective use of manual annotation, based on incremental labelling of the most ambiguous utterances, was recently proposed (Venkataraman *et al.* , 2005).

If the amount of hand-labelled data is constant, the influence of the size of the search space can balance the influence of the number of features used for classification. As shown above, the search space for one-dimensional tagsets is several orders of magnitude smaller than for multi-dimensional ones. Similarly, the search space is smaller for the linearized version of a multi-dimensional tagset obtained by enumerating only the observed combinations of tags, which is the case when a joint classifier is used for a multi-dimensional tagset. Conversely, the number of features used for classification is much smaller for multi-dimensional tagsets if "disjoint" classifiers, i.e. one for each dimension, are used in combination. In other words, the joint classifier has proportionally less data to train on than the dimension-specific classifiers.

Empirical studies are thus crucial to assess the weight of the factors discussed in Sections 6.2–6.4 with respect to the tagset dimensionality problem. In most of the experiments however, the tagset is set from the beginning based on theoretical assumptions and application goals. The scores of the existing DA taggers are therefore difficult to compare due to differences in tagsets and training/test data.

The overview provided by K. Samuel (1999, page 29) shows that the accuracy of DA tagging in different experiments—with various methods, features, data, tagsets—ranges between 50 and 75%. His own method, with the 18-tag Verbmobil tagset (Jekat *et al.* , 1995) on the Verbmobil scheduling

dialogues, reaches 65% accuracy using transformation-based learning with a set of cue words as features. Stolcke et al. (2000) report 61% accuracy on ASR output, and 71% on accurate transcripts, with the 42-tag SWBD-DAMSL tagset on the Switchboard dialogues, using HMM-based dialogue models with a forward-backward decoder. These figures show that larger tagsets do not necessarily decrease tagging accuracy if more training data, or improved classifications models, are available.

Several recent results tend to show that a joint classifier performs better than a combination of dimension-specific classifiers, when data and tagset are fixed (Clark & Popescu-Belis, 2004; Lesch *et al.* , 2005b). For instance, DA tagging using separate classifiers for each dimension of MALTUS does not perform better than using a single, combined classifier (Clark & Popescu-Belis, 2004)[24]. With a baseline of 41.9% accuracy, the joint classifier scored 73.2%, compared to 70.5% reached by the combined classifiers. The authors assume that the joint classifier performs better because of dependencies between dimensions that cannot be modelled by the independent classifiers. The frequency analysis of the ICSI-MRDA data used in (Clark & Popescu-Belis, 2004) confirms that only a small fraction of all combinations of tags do occur in the training/test data (see Section 7): of the 256 possible combinations of tags, only 50 are observed on the training data, and only 22 occur more than 20 times each, covering 99.88% of the data. The empirical analyses of labelled corpora show that only a small fraction of all combinations of tags do occur, which takes us back to the argument of the beginning of this section: the joint classifier takes advantage of a much more reduced search space than the independent ones.

If confirmed by future experiments, such results show that, below asymptotic performance, a smaller search space (i.e. a tagset that is linearized by using a joint classifier) is a more effective way to increase tagging accuracy than having (proportionally) more data for each dimension of a multi-dimensional tagset. In other words, the strong dependencies between dimensions, which cannot be modelled by independent classifiers, are a more important factor than the reduction of classification features in each dimension—in a context which is far from asymptotic performance.

8 Solutions to the Dimensionality Problem

In the analyses presented above, we identified two opposing factors that influence the definition of DA tagsets. On one side, multi-dimensional tagsets appear to receive a theoretical justification from the multiplicity of functions that utterances can fulfil. On the other side, multi-dimensional tagsets generate search spaces that are several orders of magnitude higher than those

[24]With the notations in Section 4.4, the six MALTUS classifiers are: S/B/Q/H, RP/RN/RU, AT/non-AT, DO/non-DO, PO/non-PO, and RI/non-RI.

generated by one-dimensional tagsets, and large search spaces tend to decrease the accuracy of human and automatic annotation. Several other factors influence the design of a DA tagset, such as the adequacy to a domain, task and phenomena, but the impact of these factors on the accuracy of human and automatic DA tagging seems quite difficult to generalize.

8.1 The role of dialogue theories

From a theoretical point of view, the functions of utterances in dialogues are clearly multi-dimensional, as shown in Section 3. Some theories focus on particular dimensions (see 3.3–3.8), others attempt to integrate these dimensions into a common framework (3.9), while still other theories look for higher-level principles of linguistic communication that could account, in the future, for all these dimensions (3.10). Given the complexity of some of these theories, computational linguists are challenged to derive shallower, more tractable annotation models based on these theories.

The definition of a DA tagset should make reference to the theories that are most relevant to the intended use of the tagset—as is the case with DAMSL (see Section 4.1). Reference to theoretical investigations ensures that the targeted phenomena are correctly defined, and gives access to a in-depth analyses which can be useful, for instance, to select the features that are used for utterance classification.

In addition, reference to what counts as DAs in discourse or pragmatics studies should prevent the inclusion in DA tagsets of dimensions that are not related to utterance-function, such as prosody, addressee coding, humour, emotions (see Section 3.8). Such dimensions are certainly relevant to dialogue understanding in general, but they make DA tagging more difficult if they are targeted at the same time as other dimensions.

Another important reason to encourage theoretical grounding is compatibility with other tagsets that make reference to the same theories. Even if their application goals are different, tagsets that particularize the same theory are easier to compare than tagsets based on different theories, and therefore annotated resources are also easier to convert from one tagset to another.

8.2 Dimensionality of the DA tagsets used for automatic annotation

The size of the search space appeared to be, at present, one of the most important factors in DA tagging (see Section 6.3). While multi-dimensional tagsets seem preferable on theoretical grounds, they seem to be less adapted to automatic tagging since they generate larger search spaces. However, solutions that reduce search spaces are readily available.

The size of the search space is, theoretically, the product of the sizes of

the independent dimensions of the tagset, therefore "dimensions should not be multiplied beyond necessity" (Occam's razor). Moreover, it is preferable to define 'dimensions' in the theoretical sense considered here, i.e. as sets of mutually-exclusive tags accompanied by the instruction "pick (at most) one tag" (see Section 5.1). Categories of tags accompanied by the "pick as many as apply" instructions are not genuine dimensions, as shown above, since the size of the associated search space is not n but $2^n - 1$ (if the category has n tags). In the examples above, DAMSL and ICSI-MRDA have several "dimensions" of the latter kind, generating a search space of several million combinations, whereas the dimensions of MALTUS are composed of mutually-exclusive tags, thus reducing the search space to less than one thousand combinations. SWBD-DAMSL is a truly one-dimensional tagset with only 42 tags.

To reduce the search space even more, constraints across dimensions should be found whenever possible. Theoretical analyses (such as links between speech acts and adjacency pairs, see Section 3.5) as well as empirical evidence (e.g. of the type shown in Section 7), can both be used to set such constraints. In the future, the use of global theories of communication to define DAs could spot more constraints across dimensions.

Two types of constraints can be identified. The most general type, and the one that contributes most to reduce the size of search spaces, is grouping mutually-exclusive tags into separate dimensions. This is of course fully compatible with the use of multi-dimensional tagsets. However, more efficient ways to express cross-dimensional constraints should be found, as well as methods to learn them automatically from the data. The second type of constraints across dimensions consists of individual combinations of tags that cannot occur. The only way to implement it is to linearize the multi-dimensional tagset and remove the impossible combinations from the resulting enumeration, then to use a one-dimensional (joint) classifier.

8.3 The Dominant Function Approximation: identification of dominant vs. default utterance-functions

We propose a compromise between one-dimensional and multi-dimensional tagsets that lies at the interface between the tagset structure and the tagging guidelines, for human annotators or for computer programs. Our solution is based on the idea that each utterance has a *dominant* dialogue function, an hypothesis that can be tested empirically but also in terms of applicative relevance.

The Dominant Function Approximation (DFA) proposes to consider that every utterance has only one dominant function, and that its functions in the other dimensions are the default ones, which includes the possibility that all its functions are default ones. According to the DFA, automatic DA taggers should first be required to find the dominant function of each utterance, and

are allowed to assign to the other dimensions the default value.

The DFA is thus a working hypothesis which assumes that the functional description of utterances can be slightly simplified in order to facilitate manual and automatic annotation. As such, the DFA raises a number of questions. Firstly, what are its theoretical or empirical justifications? And in particular, how can it be tested? Secondly, is the DFA of any practical value? Does it facilitate automatic tagging? And how useful are the results produced under this approximation?

8.3.1 Structure of tagset and application instructions

The taxonomy of dialogue functions that constitutes a DA tagset is best structured as a series of dimensions, or sets of mutually-exclusive tags, corresponding to the individual theoretical dimensions outlined in Sections 3.3 to 3.8. In agreement with current practice, these dimensions should not try to capture all the possible perlocutionary effects of utterances (indirect speech acts or weak implicatures), but only the "surface" functions (direct speech acts or strong implicatures). In addition, the important innovation is that in each dimension, a *default* or *unmarked* function should be identified.

The main hypothesis is that each utterance has *one dominant function*, corresponding to one tag from one dimension, and that its functions in the other dimensions are the default ones. This hypothesis represents a very strong constraint on the DA tagset, as it transforms it *de facto* into a one-dimensional tagset. Indeed, the size of the search space for a multi-dimensional tagset, with the "pick one tag from each dimension" instruction, is the product of the sizes of the dimensions (number of tags in each dimension). However, the dominant/default constraint reduces this size to the *sum* of dimension sizes only, a much smaller value in most cases[25]. The tagging instruction corresponding to the constraint is "pick only one tag from only one dimension", or just "pick exactly one tag from the tagset", with the implicit consequence that the function of an utterance in all dimensions other than the dominant one is the default (unmarked) function. This tagging instruction can be applied to both hand-labelling and automatic tagging, or to the latter only, as we explain below.

To consider an example, turn-taking is often managed smoothly by implicit or intrinsic cues, so most utterances would have an unmarked role in this dimension. However, when turn-taking must be explicitly managed, then the role of the respective utterance becomes dominant (marked) in the turn-taking dimensions, and unmarked (default) in all the other ones. For instance, if the turn is taken explicitly using a question such as "May I say something?", then its dominant role is turn grabber (or floor grabber in ICSI-MRDA terminology) and its role as a question in the speech-act

[25]Compare $5 \times 6 \times 7 = 210$ with $5 + 6 + 7 = 18$.

dimension becomes unimportant. Alternatively, if the turn was taken using an utterance such as "Shut up!", then its dominant role would become 'face-threatening', which pertains to the dimension of politeness, but is also a command. The same function could have been achieved without a command, e.g. by shouting "You idiot!".

8.3.2 Empirical assessment using human annotation

An empirical assessment of the dominant/default hypothesis is best obtained by removing first this constraint for human DA annotation—thus allowing the use of one tag from each dimension—and then counting the proportion of utterances that exhibit only one non-default function (i.e. satisfy the constraint). To speed up the annotation process, the default tag in each dimension could be explicitely indicated to annotators, and they could be instructed to mark only the non-default tags for each utterance (but still at most one tag from each dimension). Our hypothesis would be confirmed if the proportion of utterances with two or more non-default tags was small, for instance not many times higher than inter-annotator agreement. This would indicate that the amount of information lost by considering only one dominant function is "acceptable" (another account of "acceptability" could be made in terms of relevance to a language technology application, as we show below).

Annotating a single function per utterance was already one of the characteristics of the SWBD-DAMSL tagset, although these functions are in fact composed of elementary DAMSL tags. The findings that prompted the definition of SWBD-DAMSL (see Section 4.2) show that in reality only a tiny fraction of all DAMSL combinations can occur, and that these combinations can be coherently clustered into an even smaller number of classes based on their *dominant functions*, as the authors state: "we did the clustering by removing the secondary caret-dimensions" (Jurafsky *et al.* , 1997, page 2). Very few composite DAMSL labels appear in the final list of SWBD-DAMSL tags. The most frequent one is 'qy^d' for declarative yes-no questions (about 1,000 utterances out of 200,000), but its multi-functionality is not very clear: 'declarative' seems more related to form than to function. A similar case could be made for the MALTUS statistics of composite labels. These findings thus apparently support the dominant/default hypothesis.

To produce a more informative type of human annotations, the annotators could be asked explicitely to pick for each utterance one or more dominant functions, as well as zero or more secondary functions (secondary default functions could still be left unmarked). The advantage is to offer the explicit option of separating non-dominant from dominant functions. Again, the empirical test of our hypothesis could be done by checking that the number of utterances that have more than one dominant function is sufficiently small. Another test is to count how many secondary functions are

not default ones; if this number is significant, the 'secondary' option should be added to our dominant/default constraint.

The ICSI-MRDA tagset could be seen as a first step towards implementing the dominant / secondary / default constraint. The tagset has a first tier composed of so-called general tags and a second tier composed of specific tags (Dhillon *et al.* , 2004, pages 5–6). The first tier is one-dimensional, and groups mutually-exclusive tags from two theoretical dimensions, namely speech acts (statements and six types of questions) and turn-taking (backchannel, floor holder, floor grabber, and initial hold). Each intelligible utterance bears one and only one tag from the first tier, and optionally, as many other tags from second tier dimensions, which are not mutually-exclusive. Empirical studies (see Section 7) show that ca. 60% of the ICSI-MRDA utterances have only one function, and that ca. 33% have one dominant and one secondary function.

As for the MALTUS tagset, if 'statement' is considered to be the default function in the general-level dimension, and 'null' the default function in each of the other dimensions, then all utterances labelled with a general tag only, or labelled as a statement plus only one specific tag, satisfy the DFA. It appears indeed that there are ca. 97% such utterances in the ICSI-MR corpus. The main exception to the DFA (2.7% of the utterances) are utterances tagged with a question and an attention-related tag, such as tag questions. This brief analysis shows that default tags (unmarked functions) can be set *a posteriori* based on frequency counts, along with more linguistic and pragmatic considerations. To derive similar figures for ICSI-MRDA, however, a precise definition of its dimensions and then of their default tags would be required.

8.3.3 Automatic tagging of dominant utterance functions

The case of automatic DA annotation is quite different from hand-labelling. Here, the interest of using only one dominant function becomes obvious, since this allows a dramatic reduction of the search space. Therefore, a DA tagset should be accompanied, for this task, by the instruction "pick only one dominant function from the whole tagset".

The evaluation of tagging accuracy by comparison with the ground truth depends on the hand-labelling guidelines that were used. If annotators were allowed to select only one dominant function per utterance, then the evaluation metric is a simple hit-or-miss accuracy. If dominant and secondary functions could be annotated, then a more nuanced evaluation metric is needed, which must also count matches with secondary functions, as the one proposed by Lesch et al. (2005a). This metric is even more useful when the automatic tagger is itself allowed to tag dominant and secondary functions.

To conclude, we suggest that the identification of the dominant utterance

function is not only more tractable than multi-dimensional tagging, but has also practical significance to language technology applications. A program that reliably understands the dominant function of an utterance would be of great value even if human utterances had sometimes (maybe infrequently) more than one non-default function. The limitation of such a program would be its focus on the literal functions of utterances, but this would not pose problems in task-oriented human-computer dialogues, though it might overlook some of the subtleties of human dialogues in a meeting summarization task.

It is also quite obvious that for utterance generation, it is acceptable to convey two dialogue functions by using two successive utterances, if the functions cannot be packed into a single one, with the risk of being slightly less efficient than a human speaker. Of course, default functions should not be made explicit in separate utterances, to avoid the explicitation of turn-management functions, as in: "My turn now. Your booking is confirmed. Your turn now."

9 Conclusion and Perspectives

This discussion has outlined a divergence between the respective advantages of multi-dimensional and one-dimensional tagsets—theoretical grounding vs. accuracy of human and automatic annotation. We have argued that DA tagsets that are inspired by dialogue theories and are accompanied by the Dominant Function Approximation get the best of both worlds: their multi-dimensional grounding ensures understandability and interoperability, while their use in computational applications as a one-dimensional tagset increases the accuracy of DA recognition.

A number of questions remain open, starting with the actual theories that can be used to design DA tagsets and the constraints that can be found across dimensions thanks to theoretical analyses. Further empirical results related to the dimensionality problem would help investigating in more detail the contrast between one-dimensional and multi-dimensional tagsets, and to improve the estimate of the margin of error of the Dominant Function Approximation.

Acknowledgments

The work presented here was supported by the Swiss National Science Foundation through the (IM)2 project on Interactive Multimodal Information Management (*www.im2.ch*), more specifically the Multimodal Dialogue Management module (*www.issco.unige.ch/projects/im2/mdm*). I would like to thank Liz Shriberg, Barbara Peskin, and Hannah Carvey for their help with the ICSI-MRDA tagset and data; Alex Clark and Jean Carletta for feed-

back on the MALTUS tagset; and Sandrine Zufferey for her comments on earlier versions of this paper.

Appendix: Analysis of dimensionality issues using an idealized model

In an attempt to compare explicitly the independent and joint classifiers, we explored a extremely simple model of classification algorithms that do not use at all features from the data: they simply estimate the probability of each tag from the training data, then they classify utterances at random following these probabilities. In the two-dimension case, $\mathcal{C}_A$ outputs tag a_1 with probability $\pi(a_1)$, tag a_2 with probability $\pi(a_2)$, etc., regardless of the properties of each utterance ($\sum_i \pi(a_i) = 1$); the same holds for $\mathcal{C}_B$. Similarly, $\mathcal{C}_{AB}$ outputs (a_i, b_j) with probability $\pi(a_i, b_j)$ for all i, j. We suppose that enough training data is available, so that the classifiers have access to the true probability of each tag, which is the relative frequency of each tag, or pair of tags, in the data.

The expected accuracy of $\mathcal{C}_A$ is $\mathbf{E}(\mathcal{C}_A) = \sum_i \pi(a_i)^2$, that is, the probability that the correct tag is a_1 and the classifier answers a_1, plus the probability the correct tag is a_2 and $\mathcal{C}_A$ answers a_2, etc. The same applies to $\mathbf{E}(\mathcal{C}_B)$. Similarly, the expected accuracy of $\mathcal{C}_{AB}$ is $\mathbf{E}(\mathcal{C}_{AB}) = \sum_{i,j} \pi(a_i, b_j)^2$ and the expected accuracy of the combined classifier $\mathcal{C}_A\mathcal{C}_B$ is $\mathbf{E}(\mathcal{C}_A\mathcal{C}_B) = \sum_{i,j} \pi(a_i)\pi(b_j)\pi(a_i, b_j)$.

If the two dimensions A and B are independent, then the two classifiers $\mathcal{C}_A\mathcal{C}_B$ and $\mathcal{C}_{AB}$ are identical since $\pi(a_i, b_j) = \pi(a_i)\pi(b_j)$ for all i, j, and their expected accuracies are identical as well. However, if A and B are not independent, the question is whether $\mathbf{E}(\mathcal{C}_{AB})$ is always higher than $\mathbf{E}(\mathcal{C}_A\mathcal{C}_B)$ since the joint classifier $\mathcal{C}_{AB}$ has access to the joint probability distribution, while the individual classifiers $\mathcal{C}_A$ and $\mathcal{C}_B$ only estimate the individual probabilities.

Unfortunately, even for this very simple model, it is not always true that $\mathbf{E}(\mathcal{C}_{AB}) > \mathbf{E}(\mathcal{C}_A\mathcal{C}_B)$. We have examined the case when A and B have only two tags each, and found that the inequality was true in "most", but not all of the cases. We found for instance the following counter-example. If $\pi(a_1, b_1) = 0.2, \pi(a_1, b_2) = 0.6, \pi(a_2, b_1) = 0$ and $\pi(a_2, b_2) = 0.2$, then the expected accuracy of the joint classifier $\mathcal{C}_{AB}$ is *slightly lower* than the expected accuracy of the two independent classifiers put together $\mathcal{C}_A\mathcal{C}_B$. However, in this example, one of the combinations of tags, namely (a_2, b_1) is never observed, and this is correctly estimated by $\mathcal{C}_{AB}$, which one would thus expect to have a serious advantage over $\mathcal{C}_A\mathcal{C}_B$. Still, $\mathbf{E}(\mathcal{C}_{12}) - \mathbf{E}(\mathcal{C}_1\mathcal{C}_2) = -0.008$.

We also attempted to use mutual information, or KL-entropy, to compare $\mathcal{C}_1\mathcal{C}_2$ and $\mathcal{C}_{12}$, since this is a measure of the statistical correlation between

T_1 and T_2 seen as random variables. Mutual information is computed as:

$$\mathbf{I}(T_1, T_2) = H(T_1) - H(T_1|T_2) = \sum_{i,j} \pi(t_i^1, t_j^2) \log \frac{\pi(t_i^1, t_j^2)}{\pi(t_i^1)\pi(t_j^2)}.$$

The formula is symmetric with respect to '1' and '2', and $\mathbf{I}(T_1, T_2) > 0$ if T_1 and T_2 are not independent, and $\mathbf{I}(T_1, T_2) = 0$ if they are.

We attempted to relate mutual information $\mathbf{I}(T_1, T_2)$ to the difference between the expected accuracies $\mathbf{E}(\mathcal{C}_{12}) - \mathbf{E}(\mathcal{C}_1\mathcal{C}_2)$, but no such relation could be found between the two formulae, despite their similar aspects. Moreover, the counter-example above shows that $\mathbf{E}(\mathcal{C}_{12}) - \mathbf{E}(\mathcal{C}_1\mathcal{C}_2)$ can be positive or (sometimes) negative, whereas $\mathbf{I}(T_1, T_2)$ is always positive.

Given that even a very crude model of a DA classifier could not lead to formal conclusions about the performance of combined dimension-dependent classifiers vs. a joint classifier for all dimensions, it seems unlikely that more complex classifier models will allow a universal conclusion on this point.

References

Allen, James F., & Core, Mark G. 1997 (Sept./Oct., 1997). *DAMSL: Dialog Act Markup in Several Layers (Draft 2.1).* Tech. rept. Multiparty Discourse Group, Discourse Research Initiative.

Allwood, Jens. 2000. An Activity Based Approach to Pragmatics. *Pages 47–80 of:* Bunt, Harry, & Black, Bill (eds), *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics.* Amsterdam: John Benjamins.

Allwood, Jens, Nivre, Joakim, & Ahlsén, Elisabeth. 1992. On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics*, **9**(1), 1–26.

Ang, Jeremy, Liu, Yang, & Shriberg, Elizabeth. 2005. Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. *In: ICASSP 2005 (International Conference on Acoustics, Speech, and Signal Processing).*

Asher, Nicholas. 2004. Discourse topic. *Theoretical Linguistics*, **30**(2-3), 163–201.

Asher, Nicholas, & Lascarides, Alex. 2001. Indirect speech acts. *Synthese*, **128**, 183–228.

Asher, Nicholas, & Lascarides, Alex. 2003. *Logics of Conversation.* Cambridge, MA: Cambridge University Press.

Austin, John L. 1962. *How to Do Things with Words.* London: Oxford University Press.

Bales, Robert F. 1950. *Interaction Process Analysis: A Method for the Study of Small Groups.* Cambridge, MA: Addison-Wesley.

Bhagat, Sonali, Carvey, Hannah, & Shriberg, Elizabeth. 2003. Automatically Generated Prosodic Cues to Lexically Ambiguous Dialog Acts in Multiparty Meetings. *In: Proceedings of ICPhS 2003 (15th International Congress of Phonetic Sciences).*

Brown, Gillian, & Yule, George. 1983. *Discourse Analysis.* Cambridge: Cambridge University Press.

Brown, Penelope, & Levinson, Stephen C. 1987. *Politeness: Some Universals in Language Use.* Cambridge: Cambridge University Press.

Bunt, Harry. 2006. Dimensions in Dialogue Act Annotation. *Pages 919–924 of: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006).*

Bunt, Harry C. 1989. Information Dialogues as communicative action in relation to partner modelling and information processing. *Pages 47–73 of:* Taylor, Martin M., Néel, Françoise, & Bouwhuis, Dominic G. (eds), *The Structure of multimodal dialogue.* Amsterdam: North-Holland.

Bunt, Harry C. 1994. Context and Dialogue Control. *THINK Quarterly*, **3**(1), 19–31.

Bunt, Harry C. 2000. Dynamic Interpretation and Dialogue Theory. *Pages 139–166 of:* Taylor, Martin M., Néel, Françoise, & Bouwhuis, Dominic G. (eds), *The Structure of Multimodal Dialogue II.* Amsterdam: John Benjamins.

Bunt, Harry C. 2005. A Framework for Dialogue Act Specification. *In: Fourth Workshop on Multimodal Semantic Representation (ACL-SIGSEM and ISO TC37/SC4).*

Carletta, Jean. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, **22**(2), 249–254.

Carletta, Jean, & Kilgour, Jonathan. 2005. The NITE XML Toolkit Meets the ICSI Meeting Corpus: Import, Annotation, and Browsing. *Pages 111–121 of:* Bengio, Samy, & Bourlard, Herv (eds), *Machine Learning for Multimodal Interaction.* LNCS 3361. Berlin/Heidelberg: Springer-Verlag.

Carletta, Jean, Isard, Amy, Isard, Stephen, Kowtko, Jacqueline C., Doherty-Sneddon, Gwyneth, & Anderson, Anne H. 1997. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, **23**(1), 13–32.

Clark, Alexander, & Popescu-Belis, Andrei. 2004. Multi-level Dialogue Act Tags. *Pages 163–170 of: Proceedings of SIGDial 2004 (5th SIGdial Workshop on Discourse and Dialogue).*

Clark, Herbert H. 1996. *Using Language*. Cambridge: Cambridge University Press.

Clark, Herbert H., & Schaefer, Edward F. 1989. Contributing to Discourse. *Cognitive Science*, **13**, 259–294.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Core, Mark G., & Allen, James F. 1997. Coding Dialogues with the DAMSL Annotation Scheme. *Pages 28–35 of:* Traum, David R. (ed), *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*. Menlo Park, CA: American Association for Artificial Intelligence.

Craggs, Richard, & McGee Wood, Mary. 2005. Evaluating Discourse and Dialogue Coding Schemes. *Computational Linguistics*, **31**(3), 289–295.

CSTAR, Consortium. 1998. *Dialogue Act Annotation*. Unpublished document, ftp://ftp.cs.cmu.edu/project/enthusiast/ cstar/current/manual.ps.

Dhillon, Rajdip, Bhagat, Sonali, Carvey, Hannah, & Shriberg, Elizabeth. 2004. *Meeting Recorder Project: Dialog Act Labeling Guide*. Technical Report TR-04-002. ICSI, Berkeley, CA.

Di Eugenio, Barbara, & Glass, Michael. 2004. The Kappa Statistic: A Second Look. *Computational Linguistics*, **30**(1), 95–101.

Di Eugenio, Barbara, Jordan, Pamela W., Moore, Johanna D., & Thomason, Richmond H. 1998. An Empirical Investigation of Proposals in Collaborative Dialogues. *Pages 325–329 of: Coling-ACL 1998*.

Di Eugenio, Barbara, Jordan, Pamela W., Thomason, Richmond H., & Moore, Johanna D. 2000. The Agreement Process: An Empirical Investigation of Human-Human Computer-Mediated Collaborative Dialogues. *International Journal of Human Computer Studies*, **53**(6), 1017–1076.

Domingos, Pedro, & Pazzani, Michael. 1996. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. *In: ICML 1996 (International Conference on Machine Learning).*

Domingos, Pedro, & Pazzani, Michael. 1997. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, **29**, 103–130.

Doran, Christine, Aberdeen, John, Damianos, Laurie, & Hirschman, Lynette. 2003. Comparing Several Aspects of Human-Computer and Human-Human Dialogues. *In:* van Kuppevelt, Jan, & Smith, Ronnie W. (eds), *Current and New Directions in Discourse and Dialogue.* Dordrecht: Kluwer Academic Publishing.

Geertzen, Jeroen, & Bunt, Harry. 2006. Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme. *Pages 126–133 of: SIGdial 2006 (7th SIGdial Workshop on Discourse and Dialogue).*

Grosz, Barbara J., & Sidner, Candace L. 1986. Attentions, Intentions and the Structure of Discourse. *Computational Linguistics*, **12**(3), 175–204.

Harris, Zellig S. 1951. *Structural linguistics.* Chicago, IL: University of Chicago Press.

Hovy, Eduard H., & Maier, Elisabeth. 1995. *Parsimonious or Profligate: How Many and Which Discourse Structure Relations?* Unpublished manuscript. Information Sciences Institute, *www.isi.edu/natural-language/people/hovy/papers/93discproc.pdf.*

Janin, Adam, Baron, Don, Edwards, Jane A., Ellis, Dan, Gelbart, David, Morgan, Nelson, Peskin, Barbara, Pfau, Thilo, Shriberg, Elizabeth, Stolcke, Andreas, & Wooters, Chuck. 2003. The ICSI Meeting Corpus. *In: Proceedings of ICASSP 2003 (IEEE International Conference on Acoustics, Speech, and Signal Processing).*

Jekat, Susanne, Klein, Alexandra, Maier, Elisabeth, Maleck, Ilona, Mast, Marion, & Quantz, Joachim. 1995 (April 1995). *Dialogue Acts in Verbmobil.* Tech. rept. Verbmobil-Report 65. Universität Hamburg, DFKI GmbH, Universität Erlangen, TU Berlin.

Jovanovic, Natasa, op den Akker, Rieks, & Nijholt, Anton. 2005. A corpus for studying addressing behavior in multi-party dialogues. *In: 6th SIGdial Workshop on Discourse and Dialogue.*

Jurafsky, Daniel. 2003. Pragmatics and Computational Linguistics. *In:* Horn, Laurence, & Ward, Gregory (eds), *Handbook of Pragmatics*. Oxford: Blackwell.

Jurafsky, Daniel, & Martin, James H. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall.

Jurafsky, Daniel, Shriberg, Elizabeth, & Biasca, Debra. 1997. *Switchboard SWBD-DAMSL shallow discourse function annotation: Coders Manual*. Technical Report 97-02, draft 13. University of Colorado, Institute of Cognitive Science.

Jurafsky, Daniel, Shriberg, Elizabeth, Fox, Barbara, & Curl, Traci. 1998. Lexical, prosodic, and syntactic cues for dialog acts. *Pages 114–120 of: Proceedings of ACL/Coling '98 Workshop on Discourse Relations and Discourse Markers*.

Klein, Dan, & Manning, Christopher D. 2003. Optimization, Maxent Models, and Conditional Estimation without Magic. *In: Tutorial delivered at HLT-NAACL 2003*.

Klein, Marion, & Soria, Claudia. 1998. Dialogue Acts. *In:* Klein et al., Marion (ed), *MATE Deliverable 1.1: Supported Coding Schemes*. MATE (Multilevel Annotation, Tools Engineering) European Project LE4-8370.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to Its Methodology*. Beverly Hills, CA: Sage Publications.

Lesch, Stephan, Kleinbauer, Thomas, & Alexandersson, Jan. 2005a. A New Metric for the Evaluation of Dialog Act Classification. *Pages 46–53 of: Dialor 2005 (9th Workshop on the semantics and pragmatics of dialogue)*.

Lesch, Stephan, Kleinbauer, Thomas, & Alexandersson, Jan. 2005b. Towards a Decent Recognition Rate for the Automatic Classification of a Multidimensional Dialogue Act Tagset. *Pages 46–53 of: 4th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

Levinson, Stephen C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, Stephen C. 1992. Activity types and language. *Pages 66–100 of:* Drew, Paul, & Heritage, John (eds), *Talk at Work: Interaction in Institutional Settings*. Cambridge: Cambridge University Press.

Levinson, Stephen C. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature.* Cambridge, MA: The MIT Press.

Lycan, William G. 2000. *Philosophy of Language: A Contemporary Introduction.* London: Routledge.

Lyons, John. 1977. *Semantics.* Cambridge: Cambridge University Press.

Lyons, John. 1981. *Language and Linguistics.* Cambridge: Cambridge University Press.

Mann, William C., & Thompson, Sandra A. 1988. Rhetorical Structure Theory: A Theory of Text Organization. *Text*, **8**(3), 243–281.

Manning, Christopher D., & Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge, MA: The MIT Press.

Marcu, Daniel. 2000. *The Theory and Practice of Discourse Parsing and Summarization.* Cambridge, MA: The MIT Press.

McGrath, J. E. 1984. *Groups: Interaction and Performance.* Englewood Cliffs, NJ: Prentice-Hall.

Moeschler, Jacques. 1989. *Modélisation du dialogue : représentation de l'inférence argumentative.* Paris: Hermès.

Moeschler, Jacques. 2002. Speech act theory and the analysis of conversations: sequencing and interpretation in pragmatic theory. *Pages 239–261 of:* Vanderveken, Daniel, & Kubo, Susumo (eds), *Essays in Speech Act Theory.* Amsterdam: John Benjamins.

Morgan, Nelson, Baron, Don, Bhagat, Sonali, Carvey, Hannah, Dhillon, Rajdip, Edwards, Jane A., Gelbart, David, Janin, Adam, Krupski, Ashley, Peskin, Barbara, Pfau, Thilo, Shriberg, Elizabeth, Stolcke, Andreas, & Wooters, Chuck. 2003. Meetings about meetings: research at ICSI on speech in multiparty conversations. *In: Proceedings of ICASSP 2003 (IEEE International Conference on Acoustics, Speech, and Signal Processing).*

Ng, Andrew Y., & Jordan, Michael I. 2001. On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. *In: NIPS 2001 (Advances in Neural Information Processing Systems).*

Popescu-Belis, Andrei. 2003 (September 2003). *Dialogue act tagsets for meeting understanding: an abstraction based on the DAMSL, Switchboard and ICSI-MR tagsets.* Tech. rept. IM2.MDM-09. University of Geneva.

Reithinger, Norbert, & Maier, Elisabeth. 1995. Utilizing Statistical Dialogue Act Processing in Verbmobil. *Pages 116–121 of: ACL 1995 (33rd Annual Meeting of the Association for Computational Linguistics).*

Rosset, Sophie, & Lamel, Lori. 2004. Automatic Detection of Dialog Acts Based on Multi-level Information. *Pages 540–543 of: ICSLP '04 (International Conference on Speech and Language Processing).*

Roulet, Eddy, Auchlin, Antoine, Moeschler, Jacques, Rubattel, Christian, & Schelling, Marianne. 1985. *L'Articulation du Discours en Français Contemporain.* Bern: Peter Lang.

Sacks, Harvey, Schegloff, Emanuel A., & Jefferson, Gail. 1978. A simplest systematics for the organization of turn taking for conversation. *Pages 1–55 of:* Schenkein, Jim (ed), *Studies in the Organization of Conversational Interaction.* New York, NY: Academic Press.

Sadek, David. 2000. Dialogue Acts are Rational Plans. *Pages 167–187 of:* Taylor, Martin M., Néel, Françoise, & Bouwhuis, Dominic G. (eds), *The Structure of Multimodal Dialogue II.* Amsterdam: John Benjamins.

Sadock, Jerrold M., & Zwicky, Arnold. 1985. Speech act distinctions in syntax. *Pages 155–196 of:* Shopen, Timothy (ed), *Language Typology and Syntactic Description.* Cambridge: Cambridge University Press.

Samuel, Ken. 1999. *Discourse Learning: An Investigation of Dialogue Act Tagging using Transformation-Based Learning.* Ph.D. thesis, University of Delaware, Department of Computer and Information Sciences.

Schegloff, Emanuel A., & Sacks, Harvey. 1973. Opening up closings. *Semiotica*, **7**(4), 289–327.

Schiffrin, Deborah. 1987. *Discourse Markers.* Cambridge: Cambridge University Press.

Schlangen, David, Lascarides, Alex, & Copestake, Ann. 2001. Resolving Underspecification using Discourse Information. *Pages 79–93 of: Proceedings of BI-DIALOG 2001 (5th Workshop on Formal Semantics and Pragmatics of Dialogue).*

Searle, John R. 1969. *Speech Acts.* Cambridge: Cambridge University Press.

Searle, John R. 1976. A Classification of Illocutionary Acts. *Language in Society*, **5**, 1–23.

Shriberg, Elizabeth, Dhillon, Raj, Bhagat, Sonali, Ang, Jeremy, & Carvey, Hannah. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. *Pages 97–100 of: Proceedings of SIGdial 2004 (5th SIGdial Workshop on Discourse and Dialogue).*

Sinclair, John McH., & Coulthard, R. Malcolm. 1975. *Towards an Analysis of Discourse.* London: Oxford University Press.

Sperber, Dan, & Wilson, Deirdre. 1986/95. *Relevance: Communication and Cognition.* Oxford: Basil Blackwell.

Stenström, Anna-Brita. 1994. *An Introduction to Spoken Interaction.* London: Longman.

Stolcke, Andreas, Ries, Klaus, Coccaro, Noah, Shriberg, Elizabeth, Bates, Rebecca, Jurafsky, Daniel, Taylor, Paul, Martin, Rachel, Van Ess-Dykema, Carol, & Meteer, Marie. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, **26**(3), 339–371.

Traum, David R. 1999. Speech Acts for Dialogue Agents. *Pages 169–201 of:* Wooldridge, Michael, & Rao, Anand (eds), *Foundations and Theories of Rational Agents.* Dordrecht: Kluwer Academic Publishers.

Traum, David R. 2000. 20 Questions for Dialogue Act Taxonomies. *Journal of Semantics*, **17**(1), 7–30.

Traum, David R., & Hinkelman, Elizabeth A. 1992. Conversation Acts in Task-Oriented Spoken Dialogue. *Computational Intelligence*, **8**(3), 575–599.

Vandervecken, Daniel. 1990. *Meaning and speech acts. Vol. 1, Principles of Language Use. Vol. 2, Formal Semantics of Success and Satisfaction.* Cambridge: Cambridge University Press.

Vapnik, Vladimir N. 1998. *Statistical Learning Theory.* New York, NY: John Wiley.

Venkataraman, Anand, Liu, Yang, Shriberg, Elizabeth, & Stolcke, Andreas. 2005. Does Active Learning Help Automatic Dialog Act Tagging in Meeting Data? *In: Eurospeech/ Interspeech 2005.*

Walker, Marilyn, & Passonneau, Rebecca J. 2001. DATE: A Dialogue Act Tagging Scheme for Evaluation of Spoken Dialogue Systems. *In: HLT 2001 (Human Language Technology).*

Wilson, Deirdre. 1998. Discourse, coherence and relevance: a reply to Rachel Giora. *Journal of Pragmatics*, **29**(1), 57–74.

Zechner, Klaus. 2002. Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres. *Computational Linguistics*, **28**(4), 447–485.