THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *12 Janvier 2016* **par :**

Antoine Venant

Structures, commitments and games in strategic conversations

JURY

Hans Kamp	Professeur, Docteur *honoris causa*, Université de Stuttgart	Rapporteur
Hans van Ditmarsch	Directeur de recherche, CNRS, LORIA	Rapporteur
Paul Egré	Chargé de recherches, Institut Jean-Nicod, ENS Paris	Examinateur
Olivier Gasquet	Professeur, Université Toulouse 3	Examinateur
Benedikt Loewe	Professeur, Université de Hamburg, Université d'Amsterdam	Examinateur
Nicholas Asher	Directeur de recherches, CNRS, IRIT	Directeur de thèse

---

École doctorale et spécialité :
MITT : Domaine STIC : Intelligence Artificielle
Unité de recherche :
IRIT (UMR 5505)
Directeur de thèse :
Nicholas Asher

Mis en page avec la classe thesul.

Remerciements

J'ai l'occasion ici de remercier tous ceux sans qui ce travail n'aurait pas pu être accompli, tous ceux qui y ont contribué, ont eu la patience de m'écouter, me lire. Tous ceux qui, de près ou de loin, m'ont supporté (dans tous les sens du terme), pour une minute ou tous les jours. Cela tombe fort bien: vous êtes si nombreux à qui je dois beaucoup. Merci donc, a tous ceux qui figurent ci-dessous, mais également à tous ceux que je n'ai pas pu citer faute de temps et d'espace.

Merci, tout d'abord, à Nicholas Asher, pour avoir dirigé cette thèse avec tant d'enthousiasme, d'expertise et d'énergie. Merci de m'avoir proposé un sujet si passionant, et laissé tant de liberté dans la façon de le traiter. Voila deux choses qui ont beaucoup compté pour moi, parmi bien d'autres.

Je voudrais aussi remercier Hans Kamp et Hans van Ditmarsch, pour avoir accepté d'être les rapporteurs de cette thèse. Je suis très honoré de l'intérêt qu'ils ont porté à mon travail. Leurs travaux respectifs revêtent une importance considérable dans mon approche de la logique et de la linguistique; merci pour leur lecture attentive. Je souhaite remercier au même titre Paul Egré, Benedikt Loewe, et Olivier Gasquet pour avoir accepté d'examiner ma thèse et de sieger dans mon Jury.

I would like to thank Hans Kamp and Hans van Ditmarsch for agreeing to review my work. I am very honored by their interest in my work. Their respective work has had a great impact on my own approach to logic and linguistics; thank you for your careful reading of the manuscript. I would also like to thank Paul Egré, Benedikt Loewe, et Olivier Gasquet, for agreeing to examine my thesis and be in my thesis committee.

Merci à Cedric Dégremont et Soumya Paul, pour leurs precieuses idées et inestimables conseils.

Merci à l'ERC (Grant n. 269427), et aux membres du projet Strategic Conversation (STAC) pour avoir respectivement aidé à présenter mes travaux en conférence, et fourni à ceux-ci un cadre stimulant et propice.

Merci à ma famille: mes soeurs, Fabienne et Pauline, que j'admire beaucoup et qui, chacune à leur façon, m'ont toujours beaucoup inspiré. Fabienne, tu sais déja que ma curiosité a été piquée au vif d'avoir vu dévoilé, le 6 janvier 2006, un cocktail étrange d'algèbre linéaire, de linguistique et d'informatique. Aujoud'hui, je crois (et j'espère) qu'elle est encore loin d'être rassasiée. Pauline, merci pour m'avoir plusieurs fois aidé à garder la tête hors de l'eau en lui faisant prendre l'air, la haut sur la montagne. Merci à mes parents pour leur soutient sans faille, sur plusieurs continents, tout au long de mes études. Merci à eux pour leur volonté farouche de m'aider pour la soutenance et d'y assister bien qu'un continent sépare l'IRIT de leur lieu de résidence. Merci à Claudette et à Jean, pour toute l'aide qu'ils m'ont apporté depuis mon arrivée sur Toulouse et pour bien d'autres choses. Merci à Vincent et Aurore, ma petite famille bretonne.

Merci à tous les amis qui ont rendus ses 3 années si agréables. À JP, pour m'avoir enseigné que, puisqu'il est possible de tourner la molette du son a fond, il faut le faire. À Camille pour ses leçon sur la pragmatique des écharpes et des cheveux. À Julien pour m'avoir fait découvrir ce groupe qui ressemble aux Pixies. À Bridou pour s'être farci la côte pavée deux fois par semaine avec moi. À Pierre (B), Juliette, Chloé, Damien, Nadine et tous les autres, pour des motifs de même nature. Merci aussi à mes amis de plus longue date, Iovi, Gael, Nil, Pierre (M), Marion, Bastien et Bastien.

Morgane, merci à toi enfin, pour avoir été là, pour ton soutien dans la tempête, et pour bien d'autres choses encore.

Résumé

Les échanges constitués d'une succession d'actes linguistiques (dialogues, discours), témoignent d'une interaction complexe entre ces différents actes. En d'autre termes, les effets d'une action linguistique dépendent de son contexte d'execution. Le sens d'un discours, par exemple, dépend non seulement du sens de chacun d'entre eux pris isolément, mais aussi de la manière dont ses énoncés constitifs sont agencés et des liens de cohérences que ceux-ci entretiennent les un avec les autres. Il est essentiel pour un modèle linguistique de rende compte de ces liens et cela soulève plusieurs questions : comment dériver ces liens, comment les représenter au sein d'une forme logique, de quelle façon au juste contribuent-il au "sens" et/ou comment influencent-ils le choix d'une action linguistique dans un contexte particulier?

Face à ces considérations, nous nous intéressons dans un premier temps (part I) à la structure de la forme logique du discours ou du dialogue : les contraintes qu'elle doit respecter, et la façon dont différentes représentations diffèrent. Nous proposons un formalisme unifié pour les représentations adoptées par différentes théories influentes, ainsi qu'une perspective nouvelle sur les types d'information qu'elles expriment. Ces contributions s'appuient sur une interprétation commune aux différents types de formes logique dans laquelle chacune exprime un ensemble de constituants, de relations, et un ensemble de manières possible de faire porter les premières sur les secondes. Cela fournit un moyen de comparer différentes formes logiques pour un même discours à un instant donné. Mais la forme logique est un objet dynamique, modifié au fil de la conversation, et le problème se pose donc de quantifier l'impact des changements provoqués par la réalisation d'une nouvelle action linguistique. Dans un cadre abstrait où nous modélisons simplement les échanges comme des séquences de coups dialogiques, munies d'une fonction d'interprétation dans un espace "sémantique" différent, nous nous intéressons aux "distance sémantique" i.e., aux distances entre séquences de coups linguistiques dont la mesure s'appuie essentiellement sur la proximité des interprétations sémantiques respectives de deux séquences. Nous proposons de telles distances et explorons différents axiomes liant "sémanticité" des distances, structure de l'espace sémantique et déroulement de la conversation.

Un second travail (parts II and III) porte sur l'interaction entre formes logique et rationalité dans la conversation, plus particuliérement dans les conversations où les intérêts des participants divergent. Notre objectif a été la construction d'une classe de modèles en théorie des jeux dans une perspective nouvelle : les conversations en tant que séquences infinies de coups linguistiques. Un agent atteint son objectif s'il peut jouer certaines séquences et perd sinon. Ces jeux apportent une caractérisation mathématique de classes d'objectifs conversationnels décrivant la forme générale qu'une conversation réussie doit prendre. De manière cruciale, on peut justifier et expliquer via des considérations sémantiques le choix des objectifs conversationnels des joueurs : en nous appuyant sur une représentation logique du sens d'une séquence de coups (par exemple, en SDRT), nous pouvons définir les contraintes linguistiques générales (rester cohérent, consistant, crédible) qui sont des conditions nécessaires à ce qu'une conversation remplisse son objectif, et décrire les préférences des agents en termes de contenu auxquels ils s'engagent. Cela nécessite cependant une sémantique adéquate. Pour l'obtenir nous employons des outils de logique modale dynamique pour définir une logique des engagements publics (imbriqués) et l'intégrer au sein de la SDRT. Celle-ci permet de représenter les déclaration des participants vis-à-vis de leur propre engagements et de ceux de leurs interlocuteurs (incluant les implicatures), en conservant ces représentations sujettes à une conséquence (et donc une conséquence) logique adaptée. Cela permet aussi un modèle de "grounding" a granularité plus fine que les approches existantes, qui demeurent cependant axiomatisable comme des cas particulier.

Abstract

Sequential linguistic exchanges rely on a complex interplay between the different linguist acts constitutive of the exchange: the effects of a linguistic action depend on its context of performance. A book, a public allocution, a chat with a friend, a debate or a bargaining session, are as many examples of such exchanges that involve sequences of utterances. It is essential that a model of linguistic exchange accounts for these links that meaningful units have with one another, which raises a certain number of challenges: how to derive them and integrate them to a logical form, how do they contribute to the “meaning", and/or influence the choice of a given linguistic act to perform next, in a given context ?

To address these questions, we first look, in part I into the structure of the logical form of discourse and dialog: what structural constraints to adopt, what impact for a choice of constraints. We review some major theories of discourse structure and propose a unified formalism to switch from one to another, and a novel understanding of the different sort of information that these theories respectively encode. Structures of different kind are all interpreted into a common space: each structure specifies a set of constituents, one of relations, and different possible sets of scopes that the former might take over the latter. This yields a way to compare logical forms for a single discourse at some point in time. But logical forms also undergo changes as conversations unfolds, and the question of whether we can quantify the impact of a new linguistic action arrises. Placing ourselves at an abstract level where we simply model exchanges as sequences of atomic moves, equipped with a interpretation function into a different, ‘semantic’, space, we study ‘semantic metrics’ i.e., distances or similarity between sequences of linguistic moves that would base the closeness of two conversations on that of their respective semantic interpretation. We sketch an ‘axiomatic map’ of possibly atractive properties that such metrics can have w.r.t. to the properties of the semantic space, define and test several metrics against these axioms.

A second body of work (parts II and III) focuses on the interaction of logical forms and rationality in conversations, more specifically strategic dialogs, where the interest of the participants diverge. We propose a game theoretic account of such conversations within a new perspective: conversation as infinite sequences of moves. An agent is successful if he can play certain sequences, otherwise he looses. These games bring a mathematical characterization of class of conversational objectives describing the "shape" that a successful conversation must take. Crucially, we can explain why a player adopts a given set of winning sequences on semantic grounds: using a logical representation of the meaning of a sequence of moves (using e.g., Segmented Discourse Representation Theory), we can formalize linguistic constraints that are generic necessary conditions on successful plays (staying coherent, consistent, credible) , and describe agents’ preferences in term of the contents that agents commit to. This requires a semantics expressive enough, we therefore define a dynamic logic of (nested) commitments and integrate it in SDRT. This allows to represent participants’ statements about the content of theirs, or their opponent’s previous moves in the dialog (including implicatures), and keep those representations subject to a sound notion of logical consequence (and hence, of consistency). This yields also a formal account of acknowledgment and grounding that is more formal and fine-grained that traditional approaches, which can be recoverable as particular cases.

Contents

List of Figures

Introduction

What does it mean, that a sentence, or a discourse, means something? This vast question has drawn the attention of philosophers, linguists and logicians since antiquity, and more recently, of computer scientists as well.

In Artificial Intelligence (AI), interest in this question arises from a desire to model language use (basically, with a long term objective of building artificial systems that use language). Modeling language use requires to understand and capture the effects of linguistic actions, and linguistic actions can be thought of as linking *via* some relation (or *attitude*), a speaker that uses a language expression and the meaning of that expression.

How to capture these relation and meaning then? At least, we need a *formal* representation of these concepts: one that allows to formulate a set of relevant questions about them, in an explicit, unambiguous, precise way, and imposes unique, proovably correct answers for a significant subset of these questions. This is a prerequisite to any *computational* model that, in addition, would provide an effective way to compute correct answers to some (decidable) questions. Hence, we need a formal language to represent both our objects of interest (expressions, meanings, attitudes) and questions about them, and a mathematical theory over that language capturing aspects of the objects' structure relevant to the questions we want the model to answer.

We need to determine, therefore, the kind of issues about meaning to address formally, the semantic representations of language expressions that we should consequently adopt, and aspects of semantic structure that we have to theorize toward this aim.

At least two aspects of meaning seem primordial in understanding language use: issues about *entailment, i.e.,* about whether a given expression's meaning is included in that of another, and issues about *rationality, i.e.,* about one's objectives in using language and how using a given expression in a given context serves these objectives. Another arguably important aspect of the structure of natural language is *compositionality, i.e.,* the role that the internal structure of linguistics expressions (in a broad sense, where an expression might be constituted of several sentences), built out of smaller pieces, plays in answering the two previous questions.

Semantics

Several proposals exist of candidate representations and structures. Distributional semantics, for instance, represents the meaning of an expression as its sets of *contexts* of use (the exact definition of which varies with the approach). Under this view, the space of semantic representations generally inherits a structure of finite-dimensional vector space, and operators from linear algebra such as a *cosine distance*. This yields models of a purely language-internal notion of meaning, successful in providing a computational account of questions regarding *e.g.*, similarity of words', or small constructions' meaning (like adjective-noun compounds). Current distributional models however, fare far less well with the matters of *entailment* and *compositionilaty* mentioned above. Plausibly, they offer a quite powerful way to look into lexical meaning but lack a way to integrate *logical forms*, *i.e.* the internal structure of semantic objects, that supports purely logical inferences.

We will focus on another view, inherited from philosophers like Leibniz, and popularized by Frege's logic of terms and subsequent developments in symbolic logic. This view defends the idea that meanings (or thoughts) can be the objects of descriptions in a logical language. One of the characteristics of this tradition in semantics, is a take on *compositionality* where an expression of the language *refers* to something, an object of a certain type depending on the expression's syntactic category: names refer to entities, predicates to properties of entities (typically thought of as the set of those entities of which the property *holds*), so that predicates and names can combine to yield a compound expression denoting a *proposition*–namely one that states that the property denoted by the predicate holds of the entity denoted by the name. Approaches to the semantics of natural language adopting such a logical view constitute the body of *formal semantics*. Often, such approaches naturally inherit a model for entailment from their underlying logical formalism, and tend to rely on compositional interpretative mechanisms.

Works in formal semantics thus give meaning an objective existence, in the logical realm. Meanings are attached to the expressions that bear them, not to locutors' particular mental states, though they may involve some contextual parameters. *Rationality*, in this picture, rests ultimately on the logical capabilities of language users: extracting the *right* meaning of an expression, and recognizing its consequences.

Pragmatics

Using language however, involves more than just bringing about expressions and meanings. Speaking is an action, it affects the world, and it does so by placing speakers, and possibly their interlocutor, in a certain relation with these meanings they produce. It is in virtue of such relations that other then respond, argue, or react in other ways. These relations between speakers and the objects they produce, are precisely the field of study of pragmatics. As a consequence, rationality in pragmatics, not only involves agents' deductive capabilities, but also their *preferences*, what they wish to achieve in speaking or listening. Conversationalists choose to perform a given linguistic action typically if the reaction they expect from others matches their preferences.

Consider for instance a typical information-seeking dialog: A wants to know at what time her train leaves. She asks B. In so doing, she *commits* herself to the meaning of her question. B, caring to help, or simply in virtue of a desire of keeping with social convention (for instance), chooses to acknowledge A's question and answer it. If, in addition, A and B have a common knowledge that B prefers A getting her train over A missing it, then, intuitively, B should answer sincerely and A should *believe* B's answer. But then again, under the same hypothethis, if B does not know a direct answer to that question, she might answer "It should be written over there", with now an effect of A infering that B does not know a direct asnwer (because, otherwise, she would have used that answer).

Hence, *rationality* yields inferences that goes beyond the litteral meaning of what is said (B in the example above did not *say* that she does not know, nor anything that, taken independently of its particular use in context, logically entails such a conclusion), also called *implicatures*. Grice introduced the study of implicatures (and the term itself). His views on the matter still underly most works on the subject, most notably the idea that conversational implicatures are not to be considered part of litteral meaning, alternative senses attached to the expressions themselves, but obtainable from a systematic computation, *via* reasoning about what was said, and why ("*Senses are not to be multiplied beyond necessity*"–Grice 1989).

Works in pragmatics have given rise to formal accounts of meaning using game-theoretic settings: meaning is defined relative to an equilibrium, as a solution concept of a game where agents exchange linguistic actions in order to maximize some contextually given utility profile. Meaning, in this picture, is relativized to agents' actions in equilibrium: as such it captures rationality in the above sense, but obviously depends strongly on the locutors' beliefs and desires, which forms the rational basis for them to adopt an equilibrium. Such a perspective on meaning, therefore, seems to oppose formal semantics' purely logical account.

Dynamics

In the present thesis, our interest lies in a problem that, we shall see, is transversal to both the opposing views of above: formal semantics' objective conception of meaning as purely logical, and the game-theoretic conception of meaning as a solution concept of a game involving locutors' beliefs and preferences.

This problem is the *dynamics* of discourse and conversations. Both involve production of linguistic expressions in a sequence. Linguistically, they go beyond the sentencial level that many works in formal semantics focus on. As for pragmatics, the kind of games used by the vast majority of approaches, *signaling games*, models only a single exchange of a linguistic action performed by a sender, and an interpretive action subsequently performed by a receiver. We shall see that this 'local' setting brings a number of difficulties, in particular in a context where a common assumption of cooperativity is, at least, dubtious: first, there is a problem in setting local utilities when the conversation is not over and subsequent linguistic actions can make the game evolve either in a good or a very bad direction. One could try to solve such a problem by, for instance, seeing the single exchange as a subgame of a more complex game, applying a principle of backward induction to find the subgame's utilities, but we will argue that this leaves the initial problem unchanged: the real struggle is an epistemic limitation where noone is able to impose, or foresee a definitive ending for the conversation.

Interestingly, dynamic theories of meaning are, by essence, theories of the semantics/pragmatics interface. At the heart of these theories' concern, we find coreference, temporal anaphora, coherence relations between utterances, questions and their answers, agreements, corrections, grounding and other kind of phenoma which are best explained considering discourse, or conversation as a coherent whole, where a given contribution is always interpreted relatively to its context of use, involving preceding linguistic moves, or seen the other way around, where a given move affects the interpretation of subsequent ones.

Among these theories, those following the views of *dynamic semantics*, like DRT (Kamp, 1981) or SDRT (Asher, 1993) adopt a position which we can think of as intermediary between the extremes we have mentioned: in SDRT for instance, so-called *coherence* relations (or *rhetorical* relations) are infered *defeasibly* between different contributions: like conversational implicatures (of which cancellability is a fundamental property) they can, under some circonstances, be cancelled. They indeed share a lot of properties and interract with implicatures (Asher, 2013). But importantly, their inference is, most of the time, not a matter of reasoning directly about agents' intentions or preferences, but dictated by conventions linking characteristic linguistic patterns (*e.g.* a sentence mentioning an event of falling, followed by one mentioning a event of pushing) to some general communicative purpose (*e.g.*, explaining a previous utterance) supporting a given pragmatic inference. In other words, pragmatic inferences are mediated by that of a particular *discourse structure*: just like objective, structural rules of interpretation govern the possible encodings of semantics into syntax, rules of *information packaging*, govern the encoding of pragmatics into the internal organization of conversations. Conceived of this way, meaning, including implicatures, or alikes, have an objective existence that keeps in line with the view of formal semantics. Still, in the dynamic picture, locutors' **decisions** on using some move, in some context, appeal to their conversational purposes and preferences; decisions arguably involves evaluating how a move's meaning, its short term and long term effects on discourse structure, something dynamic in essence, fare with these preferences. In particular, rejecting an inference has consequences, which a locutor must carefully balance: it comes with a commitment to the negation of the particular discourse structure that underlied the cancelled inference, and so with an implicit commitment to either another inferable, licenced structure, corresponding to yet another communicative purpose, or else to incoherence, linguistic incompetence, lack of cooperativity or worse on the part of the agent who produced the material that triggered the inference.

Strategic conversations and choosen approach

Grice's account of implicatures computation crucially relies on the interplay between a principle, called the *cooperative principle*: *"Make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged."* (Grice, 1989) and four *conversational maxims*:

Maxim of Quality. Make your contribution true; so do not convey what you believe false or unjustified.

Maxim of Quantity. Be as informative as required.

Maxim of Relation. Be relevant.

Maxim of Manner. Be perspicuous; so avoid obscurity and ambiguity, and strive for brevity and order.

Several formalizations of these maxims have been proposed, some approaches have also proposed refinement, or more elementary or different principles, but in the vast majority of pragmatic accounts of rationality, a version of the cooperativity principle remains central in one way or another. In game-theoretic pragmatics for instance, cooperativity translates into a requirement that utility profiles of different agents at least partially match for their actions to have any effect. In other words, most works in pragmatics assume that communication, of any form, requires agents to share an interest.

Yet, people talk even when this is, in all appearance, not the case: in so-called strategic settings, such as political debates, negociations, bargaining sessions, disputes or trials, people might have common knowledge of their diverging interest and yet still engage in conversations.

In order to not theorize such a behaviour as irrational, there are only two possibilities for the theory: either strategic settings must 'hide' a form of cooperativity of some sort (*e.g.* politness constraints, Brown and Levinson, 1978), which requires one to explain how and why it is so, or communication, at least in a sense that needs defining, can happen without cooperativity.

In light of these considerations, it takes a new importance that dynamic semantics somehow conciliates the views of formal semantics and pragmatics: meanings determined objectively, that do not depend on locutors mental state, hence not on cooperativity either, may form the ground for communication in strategic setting. Yet static semantics will not do: attaching implicatures statically to the meaning of some contribution leaves only two possibilities: either an implicature is to be included as part of the meaning of the contribution, or it is not. But, in non-cooperative settings, an implicature can either survives, get grounded by the participants, or it can get rejected or cancelled and disappear. Whether one or the other obtains depends on subsequent moves, not on the single contribution that triggers the implicature.

Choosen approach

We will address the problem of rationality in strategic conversations by developing the dynamic solution as a compromise that encompasses both an objective notion of meaning, and an account of rationality with regards to agents' preferences. Our proposed solution will rely on a logically precise account of discourse structure, that we will extend with a dynamic, commitment-based semantics, on the one hand, and on a ressort to infinite sequential games on the other hand.

This thesis

Following the above, our general concern is the dynamic meaning of discourse and conversations, in regards of the following questions:

1. At the heart of discourse and conversations dynamics, is the idea that a new linguistic action entering the context must be related in some way to some content previously introduced (Hobbs, 1979). Several influential proposals have been developed of discourse representation that includes such relations. There is, however, no general agreement on the exact meaning of these representations. Some, like SDRT are oriented toward semantics, with a model-theoretic interpretation, other like D-LTAG seem syntactically driven, and some like Rhetorical Structure Theory (Mann and Thompson, 1986) have an ambiguous status. We will aim at understanding precisely what these representations express and how they compare, both within a given formalisms and across theories.

2. More generally, we will try to work out a notion of *semantic distance* to compare different semantic representations. The motivation for this is twofold: first, question 1 above asks for a quantification of semantic differences brought by different assumptions regarding discourse relations. It is interesting to try to quantify the semantic impact of differences in the representation of elementary units as well, and, in general, to do so abstracting as much as possible over representational peculliarities. Second, the central notion around which this thesis revolves is that of dynamics in conversational meaning, and a question that arises naturally thinking of dialog and discourse as dynamic objects, driven by rational principles, is one of quantifying the 'deviation' that a move brings *vis à vis* another. If for instance, an agent wishes that the conversation takes a particular form, and she faces an unexpected move, it is interesting to quantify how far such a move takes the conversation from her ideal obective.

3. We will examine the problem of rationality in strategic settings as previously sketched. We will in particular investigate how to overcome the limitations of traditional game-theoretic approaches with a different kind of games.

4. Finally, we will ask the question of the dynamics of agents' commitments as conversation unfolds, in presence of ambiguity. We will investigate also how this dynamics interacts with discourse relations, and how it relates to the problem of grounding.

Overview of the thesis

Chapter 1 is an introduction to formal semantics in the Montagovian tradition, and to dynamic semantics through DRT and continuation-style semantics.
Chapter 2 is an introduction to theories of discourse structure. We introduce and discuss basic notions such as elementary units, complex units and coherence relations. We then review in detail two influential theories: RST and SDRT.
Chapter 3 proposes a unified language to express different discourse formalisms using a monadic second order language and its model-theory. We show a correspondence between RST and dependency trees, under some well-defined interpretations using our language, and we propose a similarity metric for comparison of different structures.
Chapter 4 proposes an axiomatic exploration of the notion of semantic metric. We give, in an abstract setting, a precise definition of the notion of semantic metric, and its implications. We propose general distances on differently structured semantic spaces and propose candidate axioms and impossibility results.
Chapter 5 examines the specifities of conversation with respect to discourses and texts, and briefly discusses two semantics theory of conversations KoS, and D-SDRT.
Chapter 6 Details the challenges raised by strategic contexts and the difficulties that difficulties for signaling games to account for these.
Chapter 7 introduces a new kind of infinite games and explores how it can address the challenges discribed in chaper 6.

Chapter 8 Links a notion of credibility first introduced in chapter 7 to public commitments. It establishes on this basis the need for a semantic theory that deals with nested commitments and ambiguous commitments, proposes a first semantics and examines its limitations.
Chapter 9 proposes a dynamic propositional logic of commitments with ambiguous moves that overcome the limitations of the logic in chapter 8. It then discusses the problem of acknowledgments and grounding.
Chapter 10 links two different versions of the semantics in chapter 9 *via* a notion of inductive, iterated acknowledgments, then restores the full power of SDRT relational moves to the dynamic logic of chapter 9.

Part I

Structures, semantics, distances

Chapter 1

Dynamic semantics

Contents

1.1 Preliminaries

1.1.1 Model-theoretic semantics and the correspondence theory of truth

An appealing idea, and one of the most common view on semantics, consists in identifying the meaning of a sentence or a discourse *s*, with the conditions at which *s* is *true*. More specifically, Tarski (Tarski, 1936) initiated a lasting tradition of works in *model-theoretic* semantics (to which the present thesis subscribes). Following Tarski, the truth-conditions of sentences of a language (which Tarski restricts to an artificial one, though the idea has later been thoroughly applied to natural language), must be defined in a *metalanguage* (in a non-circular way). Most often, the language of Zermelo-Fraenkel's set theory is used as such a metalanguage, and *truth* is then defined in set-theoretic terms, relatively to a set-theoretic construction called a *model*.

In regards to compositionality, model-theoretic semantics naturaly conforms to a Fregean compositional unpacking of truth-conditions: language expressions 'talk about' or refer to things, and these things are elements of a hierarchy of domains (sets) in the model. These domains reflect the syntactic types of expressions in the language: each domain encompasses the possible *extensions* for expressions of a given type (*i.e.* objects that expressions of this type might refer to). For instance, *entities* in the language have extensions which are *individuals* in the model; monadic predicates, such as adjectives or common nouns in natural language, have extensions which are sets of individuals. Binary predicates, such as transitive verbs in natural language, have extensions which are sets of pair of individual, and so forth. Typically, the simple application of a predicate (*e.g. snores*) to an entity (*e.g., John*) has a truth-condition, requiring that

the extension of the entity in the model is part of that of the predicate (commonly written, given a model $\mathcal{M}$, as $\mathcal{M} \vDash snores(John)$ iff $\|John\|^{\mathcal{M}} \subseteq \|snores\|^{\mathcal{M}}$).

Arguably, model-theoretic semantics constitutes a mathematical formulation of the so-called *correspondence theory of truth*, a philosophical definition of truth favored for instance by Russel, in which truth is correspondence with the world. Model theoretic semantics subscribes to this view insofar as a model can be considered as a mathematical representation of (a given perception of) the world–at least as reflected by the semantics of the language.

1.1.2 Montague semantics and compositionality

Montague (Montague, 1988), using model-theoretic techniques, achieved a very influential semantic model for natural language. Among others, the following important features, in particular, have contributed to the success of the model:

- The semantic intepretation of each linguistic unit mirrors its syntactic category.
- The semantics builds on *intensional* logic: it is expressed in terms of objects called possible worlds: models involve a set of possible worlds, and semantic interpretations of every type of linguistic items, including sentences, are relativized to a model **and** a possible world of the model. In theoretical terms, linguistic expressions (at any level) are not directly associated extensions in the model, as briefly described above, but *intensions*, formally, functions from worlds to extensions. Truth or lack thereof, therefore, is now a property of a sentence **in a possible world** of a given model.

Montague's approach has shown that natural language might be analyzed in the same way as mathematicians' and logicians' artifical languages. Achieving the first point above, in particular, implied solving difficulties relative to differences between the syntactic structures of sentences, and their logical forms: in sentences like *I ate every cookies*, the quantified phrase every cookies is a sister constituent of the verb *ate* in the verb phrase *ate every cookies*. Yet, the logical form of this sentence is that of a universally quantified statement; roughly, *everything that has the "cookie" property is such that "I" stand in the relation of "*ate*" with it*, where the quantifyier takes scope over the binary predicate denoted by the verb. In addition, Montague's solution to these problems is 'semantic in essence', and does not require syntactic transformations such as those provided through a notion of *covert movement* of the quantified noun phrase, in some syntactic accounts.

Montague's semantics can be formulated immediately in set-theoretic terms, but the most common formulations use an intermediary language such as Church's simply typed lambda calculus (Church, 1940, which can then be equipped with a model-theoretic, denotational semantics, though there are other possibilities). Roughly, the simply typed lambda calculus involves types for entities e, truth-values t, worlds (or situations) s, constants and variables which are terms of these types, and type-construction rules: for every term T of type τ' and variable x of type τ, $\lambda x T$ is a term of type $\langle \tau, \tau' \rangle$ (the type of functions from terms of type τ to terms of type τ'), and for each terms f of type $\langle \tau, \tau' \rangle$ and X of type τ, fX is of type τ' (function application). With this surfacic overview, we can sketch Montague's account. Let us drop the intensional aspect *i.e.*, the dependence to worlds in this simple exposition:

Basic expressions of a given syntactic categories are mapped, via a lexicon, to a term of the corresponding type:

- the common noun *cookie*, is mapped to a representation $[\![\text{cookie}]\!] = \lambda x\, cookie(x)$ of type $\langle e, t \rangle$ (a function from individual to truth-values, *i.e.* a property).
- The quantifier *every* is mapped to a representation $[\![\text{every}]\!] = \lambda P \lambda Q\, \forall_x P(x) \rightarrow Q(x)$ of type $\langle \langle e, t \rangle, \langle \langle e, t \rangle, t \rangle \rangle$ (generalized quantifier).

- The transitive verb *ate* is mapped to a binary predicate $[\![\text{ate}]\!] = \lambda x \lambda y\, ate(x, y)$ of type $\langle e, \langle e, t\rangle\rangle$
- The pronoun *I* is mapped to an individual $[\![\text{I}]\!] = I$ of type e.

Montague's solution to avoid syntactic movement consists in a rule of type-raising, which states that a semantic term of type τ can be identified with the set of all properties that hold of it. Importantly, this is a logical principle (identity of indiscernibles) not a syntactic manipulation[1]. It allows to substitute an individual of type e to a term representing the set of properties that hold of that individual; for instance substituting the interpretation of I, of type e, with the term $\lambda P P(I)$ of type $\langle\langle e, t\rangle, t\rangle$. The interpretation of the transitive verb *ate* is a function from entity to functions from entity to truth-values. Applying the same principle to the first argument yields, through another transformation, a little more complex due to its role as a function's argument (yet precisely described by the theory), a new interpretation $\lambda Q \lambda x\, Q(\lambda y\, ate(x, y))$, of type $\langle\langle\langle e, t\rangle, t\rangle, \langle e, t\rangle\rangle$.

It is easy to check, that applying these two changes, and given the typing rules sketched above, the term $[\![\text{I}]\!]([\![ate]\!]([\![every]\!][\![cookie]\!]))$ is well-typed and of type t, hence represents a truth-value. In order to achieve a full semantics, one needs to complete the picture, and among other things, detail the semantics of logical constants, which can be done in more than one way (basically for instance, adding constants $\bot, \top : t$ and an predicate of *true equality*, $=_\tau$: $\langle \tau, \tau, t\rangle$ for each type τ, interpreted as external equality over extensions of terms of type τ). Anyway, a sound interpretation will have to give the same interpretation to terms that are equivalent up to *β-reduction*, *i.e.*, such that one is derived from the other replacing λ abstracted variables with contents provided by functional application $((\lambda X\, T)(U) \rightsquigarrow_\beta T[U/X])$; this provides us with a picture of the semantic mechanisms sufficient for the present introduction: one can indeed check that $[\![\text{I}]\!]([\![ate]\!]([\![every]\!][\![cookie]\!]))$ β-reduces to $\forall x\, cookie(x) \rightarrow ate(I, x)$, which, at least intuitively, seems accurate.

Reintroducing worlds in the simplified picture of above, allows to further account for contextual dependencies, propositional attitudes or reported speech and intensional contexts in general: in the example we took, one can for instance make the interpretation of *I* a function from worlds to entity, taking into account parameters of the context of enunciation and allowing elements of pragmatics to enter the context.

1.2 Dynamic semantics

1.2.1 Limits of the Montagovian approach

The semantics of Montague presented in the previous section focuses on isolated sentences, and suffers well known difficulties to account for cross-sentencial phenoma. These, led to the development of *dynamic* semantics (Kamp, 1981; Heim, 1982; Groenendijk and Stokhof, 1991; Kamp and Reyle, 1993).

One of the major limitation of Montague's approach concerns the semantic treatment of anaphora. Consider the following set of (classical) examples:

(1.2.1)A man walked in. He ordered a beer.

(1.2.2)#It is not the case that no man walked in. He ordered a beer.

(1.2.3)Every farmer who owns a donkey beats it.

(1.2.4)I bought a sage plant yesterday. I bought seven others with it.

[1]This principle as been argued to overgenerate possible readings, but more recent and refined alternatives exist. Some rely for instance on powerful syntactic formalism such as combinatorial categorial grammars, conservative of the direct mapping from syntactic to sematic types.

The problem of Montague's semantics, is that discourse referents, such as the one introduced by the noun phrase *a man*, are introduced as variables bounded by quantifiers in the semantic representations. When subsequently the pronoun *he*, then refers to that referent, predications made of the pronoun must enter the scope of the quantifier: Montague's account of the first sentence in example (1.2.1) is equivalent to the first order formula $\exists x\ man(x) \wedge walk\_in(x)$. Moreover, still following Montague, the second sentence will be rendered with a free variable, as an equivalent to $\exists b\ beer(b) \wedge order(x, y)$. Bringing the two sentences togther, should, in order to account for the correct reading for this example, result in $\exists x \exists b\ man(x) \wedge walk\_in(x) \wedge beer(b) \wedge order(x, y)$. Yet, even assuming that x and y are successfuly identified, generalizing a merging operation, such that, for instance, the merging of the two former formulae yield the latter is not trivial.

This is further confirmed by the other examples of the set: example (1.2.2) shows for instance, that the merge should not be allowed if the first formula is $\neg\neg(\exists x man(x) \wedge walked\_in(x))$, yet equivalent to the representation of the first sentence of example (1.2.1). example (1.2.3) shows that a naive account of anaphora as reduplication of content will not do: the sentence as a meaning that differs from that of *Every farmer who owns a donkey beats a donkey*. Finally example (1.2.4) shows that neither can one assume a treatment of anaphora as definite descriptions in disguise: while replacing *he* with *the man that walked in* in example (1.2.1) might do the trick, replacing *it* with *the sage plant that I bought* obviously fails with example (1.2.4) (it fails a presuposition of uniqueness).

Besides anaphora, another sort of problem concerns compositionality. Consider the examples below:

(1.2.5)John stopped talking. Mary was singing.

(1.2.6)John stopped talking. Mary sang a song.

In these two examples, It seems reasonable to assume that the meaning of the first sentence is identical in both cases. Also, interpreted in isolation, the two second sentences differ only in some aspectual parameters of the eventuality of singing they introduce; they do not involve a reference to the eventuality introduced in the first sentence. Yet, the prefered reading of the first example above is one where Mary's state of singing overlaps temporally the event of John stopping talking, and the prefered reading of the second example is one where Mary's singing temporally follows that of John stopping talking. These two readings are even incompatible with one another. A Montagovian treatment two sentences, taken in isolation, cannot account for such temporal relations between evenluaties arising when the second sentence is interpreted in context of the first.

Dynamics semantics address these issues. The main change brought by dynamic semantics, is the switch from a semantics defined with static truth-condition to one defined in terms of transformation of an input context into an output one. The semantics of a sentence is defined as such a transformation, called a *context change potential*. This is common to all dynamic theories of meaning. We will focus here on *Discourse Representation Theory*(DRT, Kamp, 1981).

1.2.2 Syntax of a DRT basic fragment

DRT constructs logical forms for input sentences that are called *Discourse Representation Structures* (DRSs). Let us recall here the syntax of DRSs for a basic fragment of DRT. As we will need such an operator later, we introduce as off now, a modal conditional > in this fragment, modeling 'normal consequences' of a given proposition.

Definition 1 (Syntax of DRSs)**.** Let $\mathcal{V}$ denotes a countably infinite set of *dicourse referents*, and $\Sigma = \biguplus_{k\in\omega} \Sigma(k)$ be a first order signature, *i.e.* a disjoint union for each integer k of sets of predicate symbols $\Sigma(k)$. Symbols in $\Sigma(k)$ are said to be of *arrity* k. A *DRS* is a pair $\langle U, K\rangle$, where $U \subseteq \mathcal{V}$ is a set of referents (called the *discourse universe* of the DRS), and K is a list of *conditions* that can recursively involve sub-DRSs.

Let [] be a symbol for the empty list (nil), and :: a list constructor symbol (for convenience, expansion of the list is made toward the right), the set of *DRSs*, $\mathcal{K}$ is defined by the following grammar:

$$\mathcal{K} ::= \langle U, K \rangle$$
$$K ::= [] \mid K :: \gamma$$
$$\gamma ::= P(x_1, \ldots, x_n) \mid \neg\mathcal{K} \mid \mathcal{K}_1 \wedge \mathcal{K}_2 \mid \mathcal{K}_1 \vee \mathcal{K}_2 \mid \mathcal{K}_1 \rightarrow \mathcal{K}_2 \mid \mathcal{K}_1 > \mathcal{K}_2 \mid x = y$$

With, $U \subseteq \mathcal{V}$; for arbitrary n, P a predicate symbol of arrity n ($P \in \Sigma(n)$) and $x_1 \ldots x_n$ n discourse referents of $\mathcal{V}$, and x, y discourse referents as well.

DRSs are often represented using a 'box' notation, whith two-floors boxes: the top floor is used to list the discourse referents and the lower to list the conditions; for instance the logical form of the first sentence in example (1.2.1) can be accounted for as the DRS $\langle U_0, K_0 \rangle$ with $U_0 = \{x\}$ and $K_0 = man(x) :: walk\_in(x)$, and the second sentence (after resolution of the anaphora) can be accounted for as $\langle U_1, K_1 \rangle$ with $U_1 = \{b, y\}$ and $K_1 = \langle beer(b) :: ordered(y, b) :: y = x \rangle$. Equivalently these two DRSs can be represented as, respectively

x, b
$man(x), walked\_in(x)$

b, y
$beer(b), ordered(y, b),$ $y = x$

.

DRS, in addition have a *merge* operation $\oplus$, defined as $\langle U, K \rangle \oplus \langle U', K' \rangle = \langle U \cup U', append(K, K') \rangle$, where *append* stands for list concatenation. In example (1.2.1) with the notations of above,

x, b
$man(x), walked\_in(x)$

$\oplus$

b, y
$beer(b), ordered(y, b),$ $y = x$

$=$

x, y, b
$man(x), walked\_in(x),$ $beer(b), ordered(y, b),$ $y = x$

Universal quantification is accounted for using $\rightarrow$ and DRS nesting: the DRS for example (1.2.3), for instance, is:

x, y	$\rightarrow$	z
$farmer(x), donkey(y), owns(x, y)$		$beats(x, z), z = y$

.

The semantics descibed in the next section ensures that this representation indeed has the right interpretation.

1.2.3 Semantics

As previously mentioned, the semantics of DRSs is defined in terms of transformation of an input context, which is called a *context-change potential*. Formally, a *context*, or *state*, is represented in DRT as a pair involving a possible world w, and a partial assignment function f mapping some of the referents in $\mathcal{V}$ to individuals in a the domain of a model. Formally:

Definition 2 (Model)**.** A model $\mathcal{M}$ is a tuple $\langle D, W, \star, I(k)\rangle$, with D a set of *individuals*, W a set of worlds, $\star$ a dynamic selection function (see below), and for each $k \in \omega$, $I(k)$ an interpretation mapping triples made of an arrity-k predicate symbol, a k-tuple of individuals and a world, to truth values: $I(k) : (\Sigma(k) \times D^k \times W) \mapsto \{\top, \bot\}$.

The set of *assignments* $\mathcal{A}$ over $\mathcal{M}$ is defined as the set of partial functions $f : \mathcal{V} \hookrightarrow D$. A partial function f' *extends* f iff $\forall x \in dom(f), x \in dom(f')$ and $f'(x) = f(x)$. A *state*, is a (world,assignment), pair (w, f) with $w \in W$ and $f \in \mathcal{A}$. We let $\mathcal{S}$ denote the set of states. A *dynamic proposition* or *context-change potential* p is a relation between states $p \subseteq \mathcal{S} \times \mathcal{S}$. We let $\mathcal{P} = \wp(\mathcal{S} \times \mathcal{S})$ denote the set of dynamic propositions. To write that two states s and s' stand in relation for p we will write indifferently $(s, s') \in p$ (prefix notation) or $s\ p\ s'$ (infix notation).

Finally, the dynamic conditional assignment function $\star$ is a function mapping, at a given world, a dynamic proposition to another dynamic proposition, it is usefull to capture a notion of 'normal consequence' of a dynamic proposition (see for instance Asher and Lascarides, 2003): $\star : W \times \mathcal{P} \mapsto \mathcal{P}$.

The semantics of DRSs is provided below:

Definition 3. The semantics is defined recursively (for DRSs and for conditions alone). In what follows we assume given a model $\mathcal{M}$, and we write $[\![\cdot]\!]$ instead of $[\![\cdot]\!]^{\mathcal{M}}$.

- $(w, f)\ [\![\langle U, [\,]\rangle]\!]\ (w', f')$ iff $w = w'$ and f' extends f, $dom(f') = dom(f) \cup U$.
- $(w, f)\ [\![\langle U, K :: \gamma\rangle]\!]\ (w', f')$ iff $\exists (w'', f'')\ (w, f)\ [\![U, K]\!]\ (w'', f'')$ and $(w'', f'')\ [\![\gamma]\!]\ (w', f')$
- $(w, f)\ [\![P(x_1, \ldots, x_n)]\!]\ (w', f')$ iff $(w, f) = (w', f')$ and $I(n)(P, \langle f(x_1), \ldots f(x_n)\rangle, w) = \top$.
- $(w, f)\ [\![\neg\gamma]\!]\ (w', f')$ iff $(w, f) = (w', f')$ and there exists no (w'', f'') such that $(w, f)\ [\![K]\!]\ (w'', f'')$
- $(w, f)\ [\![K \to K']\!]\ (w', f')$ iff $(w, f) = (w', f')$ and for all (v, g) such that $(w, f)[\![K]\!](v, g)$ there exists (u, h) such that $(v, g)[\![K']\!](u, h)$.
- $(w, f)\ [\![K \vee K']\!]\ (w', f')$ iff $(w, f)\ [\![K]\!]\ (w', f')$ or $(w, f)\ [\![K']\!]\ (w', f')$.
- $(w, f)\ [\![K \wedge K']\!]\ (w', f')$ iff $\exists (w'', f'')\ (w, f)\ [\![K]\!]\ (w'', f'')$ and $(w'', f'')\ [\![K']\!]\ (w', f')$
- $(w, f)\ [\![K > K']\!]\ (w', f')$ iff $(w, f) = (w', f')$ and for all (v, g) such that $(w, f)\ \star(w, [\![K]\!])\ (v, g)$ there exists (u, h) such that $(v, g)\ [\![K']\!]\ (u, h)$.
- $(w, f)\ [\![x = y]\!]\ (w', f')$ iff $(w, f) = (w', f'), x, y \in dom(f)$ and $f(x) = f(y)$.

We see that, the discourse universe imposes to *update* assignments to encompass new referents in their domain, while conditions always leave the assignment unchanged, but filter them, letting only through those assignments that satisfies certain properties with respect to the model.

This semantics ensures the right interpretation for the DRSs of the previous section: for instance,

$$(w, f)\ \left[\!\!\left[\begin{array}{|c|}\hline x, z, b \\ \hline man(x), walked\_in(x), \\ beer(b), ordered(z, b), \\ z = x \\ \hline\end{array}\right]\!\!\right]\ (w', f')$$

ensures that f' maps x to an individual in the interpretation of the predicate *walked in* at the actual world, b to an individual who verifies the *beer* property at the actual world, and such that both stand in the

relation denoted by *ordered* at the actual world (since $f'(z) = f'(b)$ and $f'(z)$ stand with $f'(x)$ in the latter relation). This assignment ensures in particular, existence in the model of the individual to which variable are assigned, hence the DRS represents accurately accounts for the existential reading of both undefinites in example (1.2.1).

Furthermore

$$(w,f)\ \left[\!\!\left[\ \boxed{\begin{array}{c}\\ \hline \boxed{\begin{array}{c}x,y\\ \hline farmer(x), donkey(y), owns(x,y)\end{array}} \rightarrow \boxed{\begin{array}{c}z\\ \hline beats(x,z), z = y\end{array}}\end{array}}\ \right]\!\!\right]\ (w',f')$$

requires that $f = f'$ (since the discourse universe at the top level is empty), hence it is not required that the domain of f be extended by f', and there is no garantee of existence of a farmer x or donkey y in the model, in accordance with the meaning of example (1.2.3). Morover, the semantics requires that any extension of g of f which would interpret x has an individual with the *farmer* property and y has an individual with the *donkey* property and such that x and y stand in the *owning* relation, verifies also that x and y stand in the *beating* relationship (since f must be extendable to an assignment to z that is equal to f's assignment to y and that verifies the relation). Hence, we get the correct reading on that example too, where x and y are both universally quantified.

On the other hand, assuming the following DRS for example (1.2.2) accounts for the failure of the pronoun to find a referent:

$$\boxed{\begin{array}{c}b,z\\ \hline \neg\neg\boxed{\begin{array}{c}x\\ \hline man(x), walked\_in(x)\end{array}}, beer(b), ordered(z,b), z = x\end{array}}\ .$$

We see from the semantics of the condition $z = x$, the only contexts (w,f) that can be let through the above DRS are those such that f already assigns a content to x, which means that a referent x must have been introduced in some other, previous contribution for the anaphora to succeed [2]. In particular, the anaphora will fail in an empty context, and even in a state where x is assigned an individual, there is no guarantee from this DRS alone, that z refers to somehting that is a man.

Dynamic consequence: Importantly, DRT captures a dynamic notion of entailment in natural language, in a form of *update and test*:

Definition 4. Let S and S' be two SDRSs, $S \vDash S'$, reads as S entails S' and is defined as:

$$\forall \mathcal{M} \forall s, s' \text{ if } s\ [\![S]\!]\ \text{ then } \exists s''\ (w,f)\ [\![S]\!]^{\mathcal{M}}\ (w',f')$$

Also, static Tarskian truth and validity are easilly recovered, defining: define $\mathcal{M}, s \vDash \varphi$ as $\exists s'\ s\ [\![S]\!]^{\mathcal{M}}\ s'$ (truth w.r.t. to a state), and $\vDash \varphi$ iff $\forall \mathcal{M} \forall s\ \mathcal{M}, s \vDash \varphi$ (validity). Indeed, one can define a translation t from DRS to first order formulae over the same signature, such that $\mathcal{M}, (w,f) \vDash \varphi$ iff $\mathcal{M}, f \vDash_{f.o.} t(\varphi)(w)$.

[2] We consider here for simplicity, a more DPL-oriented version of DRT, the original DRT has more constraints on referent names

Accessibility conditions: the above considerations illustrates how the semantics of DRT imposes *accessibility* conditions on referents.

A referent x is *immediately* accessible for anaphora in a DRS S iff x is introduced in the universe of a DRS S' such that S is a sub-DRS of S', or x is introduced in a DRS S' such that $S' \to S$ is a condition in a bigger DRS. *Accessibility* is the transitive closure of immediate accessibility. (Accessibility of x in S implies for instance that $S \vDash \boxed{\begin{array}{|c|}\hline z \\ \hline z = x \\ \hline \end{array}}$).

Let us conclude here, with a brief word on propositional attitude: we did not include other modal operators than the 'normal consequence' conditional > which will use in subsequent sections, but it is clear that states as world assignment leaves all liberty to introduce other modal constructions, *e.g.* a belief Bel operator, and its counterpart as a relation between worlds in the model. DRT can thus handle intensional contexts.

1.2.4 Constructing DRSs

So far we have presented DRSs and their interpretation, but not the process in which they are built from an input discourse. To this aim Kamp and Reyle (1993) develops a *construction algorithm*, inductively transforming an input syntactic derivation tree into a DRS.

This construction process, and the fact that the formulation of DRT is essentially first-order, departs from the Montagovian view, which has motivated several accounts proposing to embed DRT in a Montagovian framework, for instance λ-DRT (Muskens, 1996) and continuation-style passing (de Groote, 2006).

We briefly expose in this section the principles underlying the continuation-style passing solution of de Groote (2006). This choice has two motivations: first, it relies on the simply typed lambda calculus, which we have already briefly exposed in section 1.1.2, and importantly, does not require us to introduce a detailed account of syntax, which is convenient as the remainder of the thesis is mostly focused on semantic issues. Second, it can express Segmented Discourse Representation Theory as well (Asher and Pogodalla, 2011b, see also section 2.3 for an introduction to SDRT), a framework that underlies an important part of our contributions in the present thesis.

We illustrate this method on the treatment of an example in the spirit of examples (1.2.5) and (1.2.6), using the occasion to sketch the way dynamic semantics deals with temporal anaphora.

Let us therefeore, once again, consider the framework of the simply typed lambda calculus. This time, we take basic types to involve, at least, entities e, truth values t, and time and eventualities ϵ (we adopt a Davidsonian view with events as individuals, as is classic in DRT. We also identify here the type of times and eventualities for simplicity, as it suffices for our present needs). We will let $\Cap, \subseteq, \sqsubset$ be constants of type $\langle \epsilon, \langle \epsilon, t \rangle \rangle$ representing respectively, temporal overlap, inclusion, and precedence.

Continuation-style semantics represents a discourse as a $\lambda-$abstraction of type $[\![\Gamma]\!] = \langle \gamma, \langle \langle \gamma, t \rangle, t \rangle \rangle$ where γ is the type of input contexts. A discourse thus asks for

i) an input context i of type γ containing the effects of processing the previous discourse,

ii) A continuation o of type $\langle \gamma, t \rangle$ representing the discourse to come,

and constructs from these a truth value (type t). hence, a discourse denotes a relation between input contexts and output contexts. This is indeed what dynamics semantics pictures.

As in section 1.1.2 we will not work out a model theory, though this can be done, but rather focus on the reduction steps of the λ-calculus to obtain a normal form specifying the content of the discourse as a first-order formula, with two remaining lambda abstractions abstracting the input and output contexts.

To account for examples (1.2.5) and (1.2.6), we need to assume contexts of type γ consisting of a record of

- First, two ordered lists of accessible referents of respectively, entity and eventualities or times to account for anaphora.

- Second, a function associating to the elements of the previous list a set of attributes: for instance, genre and number for entities, tense and aspect for eventualities (*e.g.*, states/event distinction, present or past tense). Using DRT syntactic notation, we write for instance `STAT=-, TENSE=past` the features of an eventuality introduced by a verb with a stative aspect and a past tense. We let τ denote a type for such surfacic attributes (τ is a finite enumerative type).

All of this is within the representation power of the simply typed lambda calculus. To manipulate the context, we need to further assume:

1. A *rule for pragmatic inferences* $r : \langle \gamma, t \rangle$ implementing the set of rules that DRT adopts in its construction process to deal with, *e.g.* temporal anaphora (see Kamp and Reyle, 1993). To get an idea of how this works and what it does, consider as an example, that s is the last enventuality that entered the context i, that s is associated in i the feature `STAT=+` (s was introduced by a stative verb), and that the reference point in i is e'. Following DRT's account, state should include temporally their reference point, therefore $r(i)$ reduces to $e' \subseteq s$.

2. For $l \in \{e, \epsilon\}$, update functions $: :_l : \langle l, \langle \gamma, \gamma \rangle \rangle$, outputs a context updated with the add of a referent of the corresponding type l to the list of accessible referents of the input context.

3. For $l \in \{e, \epsilon\}$, $sel_l : \gamma \to l$ that selects a referent of type l in the context for coreference.

4. For $l \in \{e, \epsilon\}$, a function $v_l : \langle \gamma, \langle \tau, \langle \epsilon, \gamma \rangle \rangle \rangle$ outputs a context updated adding, if not already present, a referent of type l to the list of accessible referent **and** associating this referent with the feature passed as argument.

The last ingredient needed to put continuation-style semantics at work is a so-called *binder rule* to combine a discourse with a sentence. We can for instance use the one below:

Let $s : \gamma$ be a sentence-typed object, define $\bar{s} = \lambda i o\, s(\lambda j\, r(j) \wedge o(j))$ *i.e.* $\bar{s}$ simply transforms s in order to apply pragmatic inferences (for instance integrate anaphoric temporal condition) before applying the continuation o. The binder rule is simply provided by:

$$[\![D \cdot S]\!] = \lambda i o\, [\![D]\!] i (\lambda i'\, [\![\bar{S}]\!] i' o)$$

This binder rule states, that interpreting a sentence S in the context of the discourse D consists in feeding the continuation $\bar{s}$ to D, and that this continuation involves, interpreting s, then resolving the temporal anahpora in s using r.

If all of this is given, we can account for examples (1.2.5) and (1.2.6). To illustrate both entities and temporal anaphora let us consider this alternative example:

(1.2.7)John fell. He was running.

We can represent the lexicon as follows (For simplicity, we integrate here DRT rule for the introduction of eventuality and time, for respectively past, stative and past,non stative sentence, directly in the meaning of *fell* and *was running*):

$$[\![\text{John}]\!] = \lambda P \lambda i\, P(John)(John :: i)$$
$$[\![\text{fell}]\!] = \lambda x \lambda io\, \exists e, t\, fall(x, e) \wedge e \subseteq t \wedge t < now \wedge o(v_\epsilon(i, \texttt{STAT=-}, e))$$
$$[\![\text{he}]\!] = \lambda P \lambda i\, P(sel_e(i))$$
$$[\![\text{was running}]\!] = \lambda x\, \lambda io\, \exists s, t'\, be\_running(x, s) \wedge s \Cap t' \wedge t' < now \wedge o(v_\epsilon(i, \texttt{STAT=+}, s))$$

We leave up to the reader to check the details of the sentential derivation for both sentences. Let s_1 and s_2 respectively denote the first and second sentences of example (1.2.7). Assuming a syntactic parse gives us $[\![s_1]\!] = [\![\text{John}]\!]([\![\text{fell}]\!])$, we have after some reduction steps:

$$[\![s_1]\!] = \lambda io\, \exists e, t\, fell(John, e) \wedge e \subseteq t \wedge t < now \wedge o(John :: _e\, [v_\epsilon(i, \texttt{STAT=-}, e)])$$

and similarly $[\![s_2]\!] = [\![he]\!]([\![\text{was running}]\!])$, yielding after some reduction steps:

$$[\![s_2]\!] = \lambda io\, \exists s, t'\, be\_running(sel_e(i), s) s \Cap t' \wedge t' < now \wedge o(v_\epsilon(i, \texttt{STAT=+}, s))$$

So that the binder rule will yield, after reduction again:

$$\begin{aligned}[\![s_1 \cdot s_2]\!] =& \lambda io\, \exists e, t\, fell(John, e) \wedge e \subseteq t \wedge t < now \\ & \wedge \exists s, t'\, be\_running(sel_e(John :: _e\, [v_\epsilon(i, \texttt{STAT=-}, e)]), s) \wedge s \Cap t' \wedge t' < now \\ & \wedge r(v_\epsilon([John :: _e\, [v_\epsilon(i, \texttt{STAT=-}, e)]], \texttt{STAT=+}, s)) \\ & \wedge o(v_\epsilon([John :: _e\, [v_\epsilon(i, \texttt{STAT=-}, e)]], \texttt{STAT=+}, s))\end{aligned}$$

The last step that remains is the reduction of the call to the selection function and the pragmatic inferences rule. Assuming, for instance an empty input context i:

- $sel_e(John :: _e\, [v_\epsilon(i, \texttt{STAT=-}, e)])$ should reduce to $John$ which is the only entity accessible in the updated context $John :: _e\, [v_\epsilon(i, \texttt{STAT=-}, e)]$.
- $r(v_\epsilon([John :: _e\, [v_\epsilon(i, \texttt{STAT=-}, e)]], \texttt{STAT=+}, s))$, implementing DRT rules, should reduce to the condition $e \subseteq s$, since the context encodes that an event e was introduced followed by a state s, DRT takes the reference point to be the last introduced event if there is one, and DRT imposes that if the last introduced eventuality is a state, it includes temporally its reference point.

The above, then, finally reduces to:

$$\begin{aligned}[\![s_1 \cdot s_2]\!] =& \lambda io\, \exists e, t\, fell(John, e) \wedge e \subseteq t \wedge t < now \\ & \wedge \exists s, t'\, be\_running(John, s) \wedge s \Cap t' \wedge t' < now \\ & \wedge e \subseteq s \\ & \wedge o(v_\epsilon([John :: _e\, [v_\epsilon(i, \texttt{STAT=-}, e)]], \texttt{STAT=+}, s))\end{aligned}$$

If we had, instead treated a second sentence like *he got back up*, the context would have encoded a succession of event after interpreting s_2 and the call to r, following DRT rules would have reduced to the different constraint $e \sqsubset s$, with $\sqsubset$ representing temporal precedence.

de Groote (2006) gives a precise and motivated description of the type and content of entries in the lexicon for each syntactic categories, and, in particular, shows how the entry for quantifiers and the negation use the input context and continuation passing to encode DRT's accessibility rules.

1.3 The need for a rich semantics/pragmatics interface

DRT has been applied to a wide range of additional pragmatic phenomena, including, for instance, presuppositions (van der Sandt, 1992), attitudes reports (Asher, 1986; Kamp, 1990; Maier, 2010) and modal subordination (Roberts, 1989; Asher and Pogodalla, 2011a). Yet, as argued in Asher and Lascarides (2003), DRT, lacks to integrate something to complete the picture of the semantics/pragmatics interface. What is missing is a general picture of the implicit or explicit links that make discourse a *coherent* object, the links between utterances that Hobbs (1985) introduces as *coherence relations*.

We recap here some of the arguments that bring to this conclusion:

First, the theory's authors themselves acknowledge such a limitation: Kamp and Reyle (1993) provides the following example:

(1.3.1)John drank the beer. He liked it. Some of it ran down his chin.

the problem is that the rules we sketched in the previous section for temporal anaphora predict a reading where the event of the beer running down John's chin is temporally subsequent to John's drinking the beer, not simultaneous. The theory's author note that the correct reading should follow from inferring an elaboration relation between the two events. DRT however, was formulated before formal, explicit account of discourse structure emerged, which led the author to adopt simplying assumptions to the benefic of a formal, explicit account: "*We will therefore have to settle for a compromise. So we shall oversimplify and assume that the principles we stated above hold generally: Events always follow their reference point, states always include it*" (Kamp and Reyle, 1993, p. 528)

It seems also that DRT pictures of accessibility might be refined by considerations on discourse structure. Consider for instance:

(1.3.2)Max visited his sister yesterday. She has lovely hair.

(1.3.3)# Max visited his siter yesterday. Then he had a problem with his car and was late for dinner. She has lovely hair.

The second discourse above, if acceptable, is at least less coherent. Moreover accepting the reference of *She* to Max's sister seems to require a link of some kind between Max being late for dinner and Max's sister's hair being lovely. In terms of accessibility, DRT does not make a difference of treatment between these two examples, and DRT does not explicitely model lack coherence.

DRT's rules for temporal anaphora captures something of the semantic constraint imposed by coherence relations, but this is a partial account, as the following example shows:

(1.3.4)Max fell. John helped him up.

On such an example, DRT successfully adds to the interpretation that the helping event followed the falling event. But this exemple, arguably features more: a more coherent interpretation is one where Max falling caused John helping. More problematic is the following:

(1.3.5)Max fell. John pushed him.

The sequence of verbal tense in the latter exemple are the same as before, but this time, coherence favors a reading were the pushing caused the falling. This reading implies also that the pushing preceded the falling, which goes against the temporal rules of DRT, as we sketched them in the previous section.

For a more detailed motivation of integrating coherence relations to the dynamic framework, we refer the reader to Asher and Lascarides (2003).

Following these conclusions, we move in the next chapter to a review of coherence relations and different frameworks that use them in the analysis of discourse.

Chapter 2

Theories of discourse structure

Contents

2.1 Elementary Discourse Units and Coherence relations

Most, if not all, discourse theories assume discourse structure to be derived in a process involving two steps:

the first step consists in *segmenting* the input text, *i.e.* identifying the *Elementary Discourse Units* (EDUs). EDUs are "atoms" at the level of discourse, the content of which is assumed to be described at a lower level of syntactic and semantic analysis. The other step consists in identifying how EDUs are combined with one another by mean of *coherence relations* and yield more complex structures.

Discourse formalisms however differ in the way they address these two tasks. For instance, Rhetorical Structure Theory (RST, Mann and Thompson, 1987) assumes that discourse structure is obtained through a succesive application of abstract *schemas*, specifying how different communicative intentions of the writer are structurally realized, while another formalism such as (DLTAG, Forbes et al., 2001) offers a purely syntactic framework relying on a lexicalized grammar. Then again, other theories such as the linguistic discourse model (LDM, Polanyi, 1996) and Segmented Discourse Representation Theory (SDRT, Asher

and Lascarides, 2003) infer structure and relations using a mix of different information sources (discourse connective, lexical information, cohesion markers, the semantic content of discourse units, world knowledge).

Another important distinction between theories lies in the way they constrain possible structures (tree structures, graphs, dependency structures, ...). There is finally some variation in the interpretation and definition theories provide for both elementary units and coherence relations.

Before presenting some major theories of discourse, we briefly discuss these two central notions:

2.1.1 Elementary Discourse Units

What exactly an EDU is, is still subject to some debate and varies slightly with the formalism. There is however a strong agreement between theories that EDUs' informative content should introduce a single eventuality or state of affairs.

As such, EDUs differ from sentences, prosodic units, conversational turns and other linguistic units stemming from fields such as sentential syntax, semantics, prosody or pragmatics[3]–see for instance Polanyi, 1996, p. 5. The syntactic unit of the single (finite or non-finite) *clause* (roughly, a grammatical construction involving a single verbal predicate and its arguments) provides a very close approximation, as most of the time a single clause introduces a single discourse unit. Clausal units have for instance been used in the RST-Treebank (Carlson et al., 2001), a corpus annotated in the framework of rhetorical structure theory and one of the most used corpora for discourse parsing (see for instance duVerle and Prendinger, 2009; Subba and Di Eugenio, 2009). This choice intends to maximise the "balance between granularity of tagging and ability to identify units consistently on a large scale" (Carlson et al., 2001, p. 3). Clauses are also the smallest possible arguments of relations in the Penn Discourse Treebank (PDTB, Prasad et al., 2008), another widely used corpus annotated with discourse relations and their arguments. A few non-clausal exceptions are however permitted such as nominalizations, some anaphoric expressions, answers such as *yes* and *no*. SDRT and recent versions of the LDM further advocate a specific segmentation for cases such as infinitive complements (*Mary needs **to go***), parenthetics (*Guy Hosneld, **55 years old**, was ...*), detached adverbials (***For two decades,** John has been working at the library*), arguing that such small embedded or detached construction introduce their own EDU. The main reason put forward, is that these units must be allowed to be later on targeted as argument of subsequent coherence relations leaving the matrix clause appart, and/or to become part of a complex substructure subordinated to the same matrix, as a whole. Such a segmentation is intanciated for instance in the Annodis (Afantenos et al., 2012a) and Settler (Afantenos et al., 2012b) corpora.

In a nutshell, segmentation granularitity is ultimately driven by the need for coherence relations to get the right *scope* with respect to the theory's modeling goals. In particular, it is crucial that theories concerned with the semantic interpretation of discourse ensure that the semantic consequences of a relation are evaluated in the right context (to account for anaphora resolution, propositional attitudes, among other things), on the right arguments (otherwise the process results in a skewed logical interpretation). Polanyi et al. (2004) for instance includes a detailed proposal of a semantically driven, rather fine-grained segmentation, listing the corresponding syntactic constructions.

As a concluding remark, it is worth noticing that recent automatic approaches to discourse segmentation achieve high performances (with around 85% F-score or more), whether one requires continuously spanning EDUs (Subba and Di Eugenio, 2007), or permits embedded elementary units (Afantenos et al., 2010). Discourse segmentation is thus a well-understood, yet crucial step in the more general task of discourse parsing as segmentation errors naturally pass on to later steps of parsing such as the identification of ocherence relations.

[3] Allthough units such as sentence or turn taking are still relevant to the analysis of discourse and dialogue structure, they often introduce complex structure involving more than one unit

2.1.2 Coherence relations

As we have seen is the case for EDUs, the difference between the sets of coherence relations assumed by distinct theories is one of granularity more than one of nature: the two following principle are widely agreed on:

- Coherence relations induce a hierarchical structure over discourse. Most theories assume coherence relations to come in two kinds: coordinating, placing their arguments on the same level, and subordinating, placing one of their arguments above the other in the hierarchy[4]. It is generally agreed that this hierarchy affects available sites for finding antecedents for correference, projecting presuppostions and attaching discourse continuations. Most often this is modeled *via* the definition of a set of *open* and *closed* nodes in the discourse structure as, *e.g.* a *right frontier* constraint (Polanyi, 1985; Grosz and Sidner, 1986; Asher, 1993).

- Coherence relations have semantic consequences. These consequences impact for instance the temporal, spatial and thematic organization of a text (Hobbs et al., 1993b; Kehler, 2002; Asher, 1993; Lascarides and Asher, 1993; Hobbs et al., 1993b; Hitzeman et al., 1995), *inter alia*, but also propositional attitudes, and implicatures. Typically, the semantic conditions of a relation such as $\texttt{Narration}(a, b)$ in SDRT (or the RST-equivalent $\texttt{Sequence}(a, b)$) imposes that b's main eventuality be temporally subsequent to a's.

The main point of divergence regarding coherence relations lies in the cardinality of the set of relations theories adopt, which in turn depends on the specific aims of the theory.

D-LTAG for instance, as a purely syntactic theory, naturally identifies relations with discourse connectives. In the PDTB, the set of relations is based on the set of discourse connectives too, but a single connective might, depending on its context of use, introduce different subtypes of a given relation type. Hence, many additional semantic distinctions are made, and a hierarchic classification of relations is proposed. RST is one of the theories assuming the largest number of relations, discriminating between very fine-grained level of variation of the writer's communicative intentions. As an example, RST distinguishes between "volitional" and "non-volitional causes". Taking a different perspective, driven by the concern of providing a modular theory of compositional meaning, SDRT argue in favor of a more restricted set of relations, sufficient for the semantic interpretation of discourse and dialogue. Candidate relations must survive a kind of semantic Ockham's razor: two relations may be distinguished only if they introduce distinct semantic consequences. Hence SDRT does not distinguish between "volitional" and "non-volitional" causal relations, as the event-property of being "volitional" or not is **not** a semantic consequence of being in a causal relationship. It might however still be entailed or implicated by other parts of the discourse (consider *John pushed mary. She fell.* ***John is a terrible person*** v.s. *John pushed mary. She fell.* ***But it wasn't on purpose***).

There is however some agreement over the taxonomy of discourse relations —almost all recent theories include expressions that refer to relations like *Elaboration, Explanation, Result, Narration, Contrast, Attribution*. Sanders et al. (1992); Bateman and Rondhuis (1997); Benamara and Taboada (2015) discuss correspondences between different taxonomies.

We now review some of the major discourse theories. We put a special focus on SDRT, which will constitute one of our starting point for the semantic and pragmatics analysis of dialogue, and will therefore be of heavy use in all parts of the present thesis.

[4] Not all theories use the terms "subordinating" and "coordinating", and may have instead theory-specific closely related notions, such as the nucleus/satellite disctinction in RST

2.2 Rhetorical Structure Theory

2.2.1 Overview

In the words of the theory's authors, RST is "a descriptive framework for text" which "identifies hierarchic structure in text" and "describes the relations between text parts in functional terms" (Mann and Thompson, 1987). RST thus adopts a descriptive and functional perspective on discourse structure: coherence relations represent a way for an *analyst* to describe what he thinks the intended communicative function[5] of two adjacent spans of text is. The analyst's judgments are explicited by the constraints associated with the relations.

RST makes a distinction between relations involving a Nucleus and a Satellite (which we will write NS or SN-relations, depending on the relative positions of the nucleus and satellite), and relations involving two nuclei (multinuclear, or NN-relations). This distinction is quite analoguous to the distinction between subordinating (NS/SN-relations) and coordinating (NN relations), discussed in the previous section. RST's specifity however, is to attach this information to the discourse unit itself, and not to the relation. A given unit is either a nucleus or a satellite, never both, but it might contain other nuclei or satellites as subunits. Relations are associated a set of up to 4 constraints (given in the form of statements in natural language). These constraints express properties that must be met by, respectively, each of their argument, their combination, and the writer's intention behind using the relation. We provide for instance the definition of the relation `Cause` as used in the RSTTB:

Constraints on `Cause` (NS or SN-relation)

Nucleus: $\varnothing$

Satellite: $\varnothing$

Combination of N and S: The situation presented in the nucleus is the cause of the situation presented in the satellite. The cause, which is the nucleus, is the most important part. The satellite represents the result of the action.

Writer's intention: The intention of the writer is to emphasize the cause.

An analysis in the framework of RST consists in the successive application of *schemas* to the input text. A schema is an abstract specification of a (flat) fragment of discourse structure. Formally, a schema can be defined as a graph $s = (V, E)$ such that $V = \langle v_{0a_0} \dots v_{na_n} \rangle$ is a linearly ordered set of vertices (placeholder for relations' arguments), with each v_i annotated with a nucleus/satellite token $a_i \in \{N, S\}$ and $E \subseteq V \times V \times \mathcal{R}$ is a set of edges annoted with coherence relations (we let $\mathcal{R}$ denote the set of coherence relations). Furthermore, only two cases are permitted:

- Either there is a unique i such that $a_i = N$, and every edge in E links a satellite to this nucleus (single nucleus).
- Or $\forall i\, a_i = N$ and $E = \{R(v_i, v_{i+1}) \mid i < n\}$ for some multinuclear relation $R \in \mathcal{R}$ (multinuclear schema).

For instance, the schema corresponding to instances of the NS-relation `Circumstance` is simply given by $V = \langle v_{0N}, v_{1S} \rangle, E = \{\texttt{Circumstance}(v_0, v_1)\}$. *Applying* s to a span of text T simply consists in splitting T into n adjacent subspans of text $t_1 \dots t_n$, mapping them to each of to the placeholders for relations' arguments in the schema. In other words, applying s to T consists in intanciating within T the fragment of

[5] *i.e.* the intented effect that the writers think he achieves on a reader.

structure that s specifies. The application is nevertheless permitted only at the condition that it is plausible to the analyst, that the constraints associated with relation R_i hold between t_i and t_{i+1}, for each R_i. Once a schema has been applied to a span T, T and the subspans involved in the schema application become DUs.

The analysis ends when:

1. A schema has been applied to the span encompassing the whole discourse.
2. Every span used as argument to a relation is either an EDU or a schema has been applied to it.
3. No text span is used as an argument in two distinct schema applications.

The analysis results in a hierarchic structure called an RST Tree.

As appears from the above, an RST tree is tight to the point of view of the analyst who produces it. Now what an analyst thinks the writer intends to communicate might obviously differ from what the writer actually intends to communicate (let alone what he actually commits to). However, a semantically competent analyst might of course not produce just any structure, at least if he believes the writer to be semantically competent and rational as well. Intuitively, such assumptions impose that the analyst judges that the author intends to communicate, at least, what he actually commits to (by conventionl meaning). Thus, even though RST is not a theory of semantic meaning, an RST tree can be thought of as expressing both a set of constraints over the semantic interpretation of a text, and a plausible interpretation of the writer's intentions **compatible with these constraints**. Indeed Mann and Thompson (1987) insists that the analyst "has access to the text, knowledge of the context in which it was written, and shares the cultural conventions of the writer and the expected readers, but has no direct access to the writer or other readers", analyst judgments are thus assumed objectively determined by the content of the text, not by the analyst subjective beliefs. Therefore, it makes sense to refer to the set of possible RST trees **for** a given text. Anticipating a little on the next chapter, these considerations also provide a justification on a comparative approach of RST to other discourse theories based on the semantic scope of coherence relations.

We now discuss in more details the set of RST Trees that might be output by the above process, and their use in annotated corpora.

2.2.2 Structures and Corpora

In order to discuss the structural constraints governing RST Trees, let us first introduce a few more precisions about RST schemas:

RST defines five type of schemas, consisting respectively in:

1. A single NS- ($a_0 = N, a_1 = S$) or SN- ($a_0 = S, a_1 = N$) relation: $V = \langle v_{0a_0}, v_{1a_1} \rangle, E = \{R_0(v_0, v_1)\}$.
2. A single NN-relation (just like the above, except that $a_0 = a_1 = N$).
3. For some specific relations such as `List` or `Sequence`, a multinuclear schema exist for arbitrary long sequences of such relations *i.e.* for all $n \in \mathbb{N}$ RST has a schema

   $$V = \langle v_{0N} \dots v_{nN} \rangle, E = \{R(v_i, v_{i+1}) \mid i < n\}$$

   with $R =$ `Sequence` or $R =$ `List`.
4. For some pair of relations such as (`Motivation`, `Enablement`), a common nucleus linked to a satellite by each of the relations (a schema exists for each position of the nucleus). For instance with the nucleus in the middle:

   $$V = \langle x_{0S}, x_{1N}, x_{2S} \rangle, E = \{\texttt{Motivation}(x_0, x_1), \texttt{Enablement}(x_1, x_2)\}$$

5. A *Join* schema, linking two nucleus but involving no relation. This schema is merely a conventional way to represent the absence of any relation. We can therefore reduce it to the multinuclear schema, by assuming a specific multinuclear `Join` relation and a convention that this relation actually stands for the absence of other relations (which is indeed the representation adopted practice).

In the following, let $\subseteq$ denote textual inclusion of spans, and $<_t$ denote textual precedence. $<_t$ is naturally extended to spans by defining that $s_1 <_t s_2$ iff the left boundary of s_1 appears prior to the left boundary of s_2 in the text. Notice that, allthough Mann and Thompson (1987) defines a span as an "uninterrupted linear interval of text", excluding the possibility that segmenting an embedded EDU yields a discontinuous span for its matrix EDU, this definition is easily relaxed into a weaker and consensual assumption of non partially overlapping spans. Adjacency, required in RST for two spans to be linked by a coherence relation, is simply reinterpreted as immediate succession with respect to $<_t$, and nothing changes besides this. The choice of the left boundary for spans' ordering naturally makes a discountinous matrix EDU textually precede its splitting embedded EDUs, which is the convention used in corpora featuring such segmentation, as *e.g.* ANNODIS. As a consequence, segmentation raises no issue to the comparison of RST Trees and other structures proposed in the next chapter.

The denomination "RST Tree" is legitimate because DUs (spans) form a tree with respect to immediate-inclusion (s' is a child of s if $s' \subsetneq s$ and there is no span s'' such that $s' \subsetneq s'' \subsetneq s$), moreover each non-elementary DU T is itself an acyclic graph $\langle V[t_i/v_i], E\rangle$ obtained from the application of a schema $\langle V, E\rangle$ to T, replacing each v_i by T's i^{th} (w.r.t $<_t$) child. We thus see that, due to the schemas of type 3 and 4 above, an RST Tree is indeed a little more than a regular labelled tree.

As previously mentioned, RST is a theory underlying several corpora, such as the RST Treebank (RSTTB, Carlson et al., 2001) and the Potsdam commentary corpus (Stede, 2004). In practice, these corpora involve a huge majority of schemas 1 and 2, (if any occurence of 3 at all). Indeed, approaches to discourse parsing such as duVerle and Prendinger (2009) focus on structures obtainable from schema 1 and 2 only. As a consequence, when refering to RST, we will often focus our efforts on such structures, without multiple relations-schemas (especially when relating RST Trees and DGs in chapter 3). It is straightforward to see that, for this class of simpler RST Tree, each DU's internal structure involves a single edge between two children DUs, and therefore is fully describable as a binary trees whith each node labelled with both nuclearity (except for the root) and a single relation label:

Definition 5 (Binary RST Tree)**.**

- An EDU is an RST Tree.
- If R is a nucleus-statellite relation symbol, s_1 and s_2 are both binary RST Trees the root of which are contiguous spans, and $\langle a_1, a_2\rangle \in \{\langle N, S\rangle; \langle S, N\rangle\}$ then $R(t_{1a_1}, t_{2a_2})$ is an RST Tree.

2.2.3 Examples

To conclude the present introduction to RST, we provide two examples and their RST structures:

(2.2.1) [Interprovincial Pipe Line Co. said]$_{C_1}$ [it will delay a proposed two-step, 830 million dollar [(US $705.6 million)]$_{C_3}$ expansion of its system]$_{C_2}$ [because Canada's output of crude oil is shrinking.]$_{C_4}$ (from RSTTB, wsj_2363)

(2.2.2) ["he was a very aggressive firefighter.]$_{C_1}$ [he loved the work he was in,"]$_{C_2}$ [said acting fire chief Lary Garcia.]$_{C_3}$. ["He couldn't be bested in terms of his willingness and his ability to do something to help you survive."]$_{C_4}$ (from Egg and Redeker, 2010)

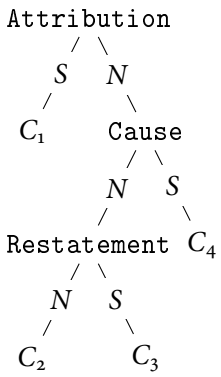
The segmentation can be read from the examples with elementary discourse units placed in square brackets, indexed with the unit's name. Note that example (2.2.1) features an embedded unit C_3 which splits C_2 into two discontinuous spans. RSTTB has actually no units with discontinuous spans, instead a

more complex encoding is used where discontinuous units are segmented as two different continuous units, linked by an "artificial" `Same_Unit` relation. We drop this here, simply taking C_3 has having a discontinuous span. We will generally consider `Same_Unit` (or its pendant `Fusion` in ANNODIS) as an empirical way to recover from segmentation mistakes.

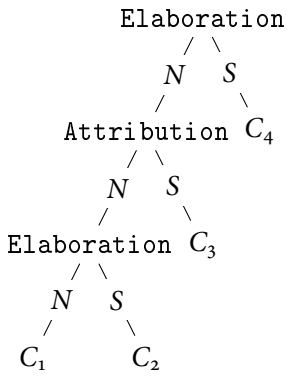
Figure 2.1 graphically depicts the (binary) RST Trees for exemples examples (2.2.1) and (2.2.2). To improve lisibility, the N/S annotations of a child node is written midway on the edge to its parent node.

Figure 2.1: RST Trees for examples (2.2.1) and (2.2.2)

(a) RST Tree for example (2.2.1)



(b) RST Tree for example (2.2.2)



2.2.4 Summary

We have seen that RST proposes to model the hierarchic organization of discourse with tree-shaped structures. Nodes of these structures correspond to spans of text, labelled with coherence relations holding between different (adjacent) subspans. Whether or not a relation holds is something that the analyst must judge, given the constraints associated with each relation (formulated in natural language). RST's take on discourse structure is to be understood in terms of *communicative function* of spans *vis à vis* each other.

Although the interpretation of RST trees, in some way, tells us something about the semantic interpretation of discourse, RST differ from a formal theory of semantic meaning, such as SDRT which we introduce in the next section. As we will see, one of the main switch in perspective, is the vision of relations as logical operators. A consequence is, for instance, that SDRT emphasizes the impact of the semantic scope of a coherence relation and precises the compositional semantic interpretation of complex units. Such notions are less important to RST's descriptive perspective, where the exact contribution of each subunit of a complex DU to the satisfaction of a relation's constraints is consequently a little blurred.

2.3 Segmented Discourse Representation Theory

2.3.1 Overview

Segmented Discourse Representation Theory (SDRT) Asher (1993); Asher and Lascarides (2003), is a semantic theory which inherits a framework from dynamic semantics and enriches it with coherence relations.

SDRT, as DRT, is a representational theory: it assumes the existence of *logical forms*, structured mental representations obtaining at a level intermediary between the surface structure of discourse and semantic values. The main insight of SDRT is to adopt a set of logical forms of which discourse structure is made a

first class citizen. In these logical forms called Segmented Discourse Representation Structures (SDRSs), coherence relations play the role of discourse-level predicates scoping over lower-level discourse constituents. these consitutents may themselves recursively exhibit an analog structure. Like DRSs in DRT, SDRSs are incrementally refined as discourse unfolds, and are equipped with a model-theoretic interpretation, providing at each step a transition from internal representations to semantic values. The set of SDRSs and their model-theoretic interpretation together constitute the *logic of information content*.

An SDRS roughly corresponds to a set of *discourse labels*, one of which is the *top* label. Labels are assigned a content, which is

- Either a semantic representation of an EDU, in a given *base-level language*, such as DRT or DPL.
- Either a set of subordinate labels linked by discourse relations. (Complex Discourse Unit, CDU).

The model-theoretic interpretation of a CDU is obtained by applying the semantic consequences associated with coherence relations to the recursively unpacked semantic content of the relations' arguments. The semantic interpretation of an EDU label is simply the one of its content, furnished by a *base-level* dynamic logic. The interpretation of the SDRS is by definition that of its top-label. We will detail this in the next subsection.

An important specifity of the theory is the distinction made between *infering* and *evaluating* SDRSs. Asher and Lascarides argue that infering the structure of a given discourse is computationally less challenging than deciding a property of the structure's interpretation. Consider for instance the following discourse: *[John is a brilliant mathematician]*$_1$, *[he found a polynomial algorithm deciding 3-SAT.]*$_2$ *[He concluded that P=NP]*$_3$. On the hone hand, it seems rather easy to infer that the rhetorical function of 2 and 3 is to support the claim made in 1 (in SDRT terms, `Elaboration` and `Explanation`* can be inferred), even ignoring the precise denotation of terms such as 3-SAT, P, or NP. It is as easy to spot that 3 rhetorically comes as a result of 2 (plausibly triggered by the joint presence of the lexical items *found* and *concluded*). On the other hand, actually evaluating whether the result between 2 and 3 obtains would require at least some familiarity with John and basic complexity theory. Similarly, evaluating the truth of 2 on its own would require to know John's algorithm (if indeed there is one), as well as expert knowledge of theoretical computer science.

Following this line of reasoning, the logic of information content is designed to be (at least) first order expressive, with an undecidable (indeed non recursively enumerable) consistency check problem, whereas the inference of discourse structure, assumed to be a decidable problem, is internally modeled by mean of a *glue logic* (also refered to as the *logic of information packaging*). The glue logic is intended to have a restricted access to SDRSs' content and world knowledge, to be kept decidable, while remaining sufficiently expressive to model structural inferences. Indeed the glue logic's domain objects are (syntactic) fragment of SDRSs, not world's individual or abstract semantic entities. SDRT thus aim at separating *information content* from *information packaging* and explain the flow of information between the two layers.

This concern for separating the intepreter's evaluation of the discourse (what he believes) and the discourse content ("what is said", derived on linguistic grounds) is central to the perspective adopted by SDRT, and distinguishes the theory from other closely related semantic-oriented models such as, for instance, Hobbs et al. (1993a)'s abductive theory where every step of interpretation is conducted in the same first order object language. Such a distinction further applies to other *belief-change* models of discourse interpretation, building on Searle (1969) and Grice (1975) vision of meaning and speech acts, such as Grosz and Sidner (1986). A detailed review is presented in Asher and Lascarides (2003, chapter 3). The general claim is that, even though the semantic interpretation of discourse may affect an interpreter's cognitive state, semantic understanding **cannot be equated** with such cognitive changes. We adopt in the present thesis such a position, which constitute the starting point for our analysis of strategic conversation through the game-theoretic proposal in chapter 7 and the commitments-based semantics developped in chapters 8 to 10.

We now examine in more details SDRSs and their interpretation.

2.3.2 The syntax of SDRSs: directed acyclic graphs

SDRT assumes provided a base-level dynamic logic $\mathcal{L}_0$, equipped with a class of models, and for each model $\mathcal{M}$, a set of *states* $X^{\mathcal{M}}$ and (dynamic) interpretation $[\![\cdot]\!]^{\mathcal{M}} : \mathcal{L}_0 \mapsto X^{\mathcal{M}} \times X^{\mathcal{M}}$. To ease notational clutter, we will drop the $\cdot^{\mathcal{M}}$ exponent notation whenever the model is clear from context. for $x, x' \in X^{\mathcal{M}}$ and $\varphi \in \mathcal{L}_0$, we will use infix notation $x[\![\varphi]\!]^{\mathcal{M}}x'$ with the same meaning as $\langle x, x' \rangle \in [\![\varphi]\!]^{\mathcal{M}}$.

Let Π be an infinite countable set of discourse labels, and $\mathcal{R}$ be a set of rhetorical relations. Let $\mathcal{R}_{\leftarrow}$ and $\mathcal{R}_{\rightarrow}$ denote two subsets of $\mathcal{R}$, standing for, respectively *left veridical* and *right veridical* relations. Anticipating a bit on the semantic consequences of relations, left (resp. right) veridical relations are those which semantically imply the content of their left (resp. right) argument. A relation is *veridical* if it is both left and right veridical, and *non-veridical* if it is neither left nor right veridical. Veridical relations include for instance `Elaboration`, `Explanation`, `Result`, `Contrast`, `Parallel`. Examples of left veridical, non right veridical relations are `Attribution` and `Goal`. Examples of non veridical relations are `Alternation` and `Conditional`. Finally, assume a partition of $\mathcal{R}$ into coordinating, and subordinating relations.

We are now ready to define the syntax of SDRSs. Looking ahead to chapter 3 we provide a reworked, graph-theoretic version of the definition in Asher and Lascarides (2003), slightly more complex, but with the advantage of making all structural constraints on SDRSs fully explicit at once (also integrating in the process considerations on discourse structure such that the ones presented in Asher et al. (2011)).

Definition 6 (SDRS)**.** An SDRS is a tuple $S = \langle A, O, E, \pi_{\text{top}}, f \rangle$ where

- A is a finite subset of Π (set of discourse labels)
- $\pi_{\text{top}} \in A$
- $O \subseteq A \times A$ is a set of edges such that $\langle A, O \rangle$ is a tree rooted in π_{top} (in the graph-theoretic sense[6]). Edges in O need not be directed, but direction is implicit and can always be retrieved as follows: define the "oustscope" order $\prec^*$ on labels: $\pi_2 \prec^* \pi_1$ iff $\pi_1 \in \mathcal{P}(\pi_{\text{top}}, \pi_2)$, where $\mathcal{P}(\pi_{\text{top}}, \pi)$ denotes the (unique) path from π_{top} to π in $\langle A, O \rangle$. Finally direct edges from left to right in descending order w.r.t. $\prec^*$.
- $E \subseteq \mathcal{R} \times A \times A$ is a set of directed edges labelled with discourse relations, such that $\langle A, E \cup O \rangle$ is a directed acyclic graph (with O directed as above), and such that O is closed for incoming coherence relations, *i.e.* whenever we have $\langle \pi_0, \pi \rangle \in O$ and $R(\pi_1, \pi) \in E$, we have $\langle \pi_0, \pi_1 \rangle \in O$ (as we did when reviewing RST, we write $R(\pi, \pi') \in E$ with the meaning of $\langle R, \pi, \pi' \rangle \in E$). A final constraint is that two labels might not be linked by both a coordinating and a subordinating relation.
- f is a function mapping EDU labels to formulae in $\mathcal{L}_0$. An EDU label is a label minimal for $\prec^*$. Any other label is a *complex discourse unit* (CDU) label.
- S respects the *right frontier constraint* the definition of which will be provided after introducing a few additional notations.

Put more simply, an SDRS is essentially an acyclic directed graph with a "top" label, and two kind of edges: edges in O describe the subunits of complex discourse units, while edges in E encode the coherence relations holding between different discourse units. Labels at the bottom of the structure represent EDUs, and are associated the corresponding representations in the lower language *via* f.

[6] *i.e.*, a graph such that there is a unique path from π_{top} to any other node. Equivalently we could have dropped π_{top} from the definition of an SDRS, and defined O as a set of directed edges.

Notice that, while the definition explicitely forbids an incoming coherence relation to "penetrate" a complex discourse unit, it licences outgoing coherence relations to cross the boundaries of CDUs: it is possible that an edge in E links π_1 to π_2 when, for instance, π_1 is a child of π and π_2 is a sibling of π. This is a desirable feature of the theory, and allows the modelling of *discouse asides* (*e.g.* digressions or commentary performed while in a embedded, subordinated discourse topic).

The following definition permits to retrieve the recursive structure of discourse from the graph-theoretic definition: define $\pi \prec \pi'$ as $O(\pi, \pi') \in O$, or equivalently: $\pi \prec^* \pi'$ and $\neg\exists\pi'' \notin \{\pi, \pi'\}\ \pi \prec^* \pi'' \prec^* \pi'$. Note that $\prec^*$ is the transitive relexive closure of $\prec$, which justifies the notational choice. $\pi \prec \pi'$ reads as "π' immediately outscope π".

Since EDU labels represent elementary units, they inherit the textual ordering of EDUs, which we let $<_T$ denote. This is the last ingredient to the definition of the Right Frontier and associated constraint:

Definition 7 (Right Frontier)**.** The right frontier of an SDRS S is the set of nodes $\mathrm{RF}(S)$ inductively defined as the least set such that:

- $\pi_{\text{last}} \in \mathrm{RF}(S)$, where π_{last} is the last EDU label in textual order (*i.e.* w.r.t. $<_T$)
- $\forall \pi \in \mathrm{RF}(S) \forall \pi' > \pi\ \pi' \in \mathrm{RF}(S)$
- $\forall \pi \in \mathrm{RF}(S) \forall \pi'$ s.t. $R(\pi', \pi) \in E$ for some **subordinating** R, $\pi' \in \mathrm{RF}(S)$.

Definition 8 (Right Frontier constraint)**.** The definition is recursive: a single EDU label π verifies the right frontier constraint. For the recursive step, let $S = \langle A, O, E, \pi_{\text{top}} \rangle$ (omitting the assignation f for convenience) and let $S - \pi_{\text{last}}$ denotes the minor of S obtained by the following process: remove π_{last} from A as well as any edge in E or O involving π_{last}. If the parent π' of π_{last} has no other descendant, π' is by definition an EDU label of $S - \pi_{\text{last}}$, marked as last in textual order.

S verifies the right frontier constraint iff:

- $S - \pi_{\text{last}}$ recursively verifies the right frontier constraint.
- $\forall R(\pi', \pi) \in E$, $\pi' \in \mathrm{RF}(S - \pi_{\text{last}})$.

This definition expresses that the incremental construction of S must be performed in such a way that new information is always attached on a node of the right frontier of the current structure, or otherwise starts a new complex discourse unit which itself has to attach on the right frontier. The right frontier constraint, thus model possible sites for attachment. As we will see later, it also models possible antecedents for anaphora resolution.

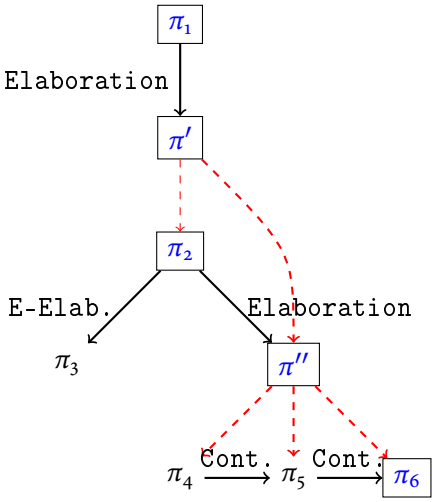
To illustrate the above definitions, we provide on fig. 2.2 an example of discourse and associated SDRS (of course it remains to explain how SDRT predicts this structure to be interpreted and infered. Graphical convention are as follows: edges in E (coherence relations) are drawn in black, horizontaly when coordinating, vertically when subordinating (subordinating edges may be inclined for a better rendering). Edges in O (Labels outscoping relations) are drawn dashed and in red. Nodes of the right frontier are squared in black, with blue labels. The example involves relations of elaboration, entity elaboration (`E-Elab.`) and continuation (`Cont.`). To avoid overloading the drawing with too much edges, the top label is not represented, but it is implicitely understood that it outscopes all other labels (and immediately outscopes π, π' and π_3).

The example involves an instance of discourse aside: we see that the relation of entity elaboration linking π_3 to π_2, "escapes" the CDU label π' and thus forms a discourse aside at the level of the top-label.

We now turn to the interpretation of SDRSs.

Figure 2.2: Example discourse (translated from ANNODIS) and associated SDRS

(2.3.1)[Principles of natural selection.]$_{\pi_1}$ [The theory of natural selection [as it was initially described by Charles Darwin,]$_{\pi_3}$ rest on three principles:]$_{\pi_2}$ [1. the principle of variation]$_{\pi_4}$ [2. the principle of adaptation]$_{\pi_5}$ [3. the principle of heredity]$_{\pi_6}$



2.3.3 The logic of information content

The interpretation of SDRSs builds on, and extends the dynamic semantics $[\![\cdot]\!]$ of the base-level language. Hence, the semantics of an SDRS is understood as a context-change potential . The idea is that, at each level of the hierarchy, the interpretation of a complex unit is obtained by applying the semantic consequences of relations to their arguments, the semantics of which is recursively unpacked.

Since dynamic cunjunction is not commutative (for instance, a cunjunct might introduce a new referent into the context which a subsequent cunjunct might then pick up), we must see that coherence relations and their arguments are evaluated in the appropriate order. Basically, this amounts, when a CDU label involves more than one relation, to evaluate the relations in an order that respects the textual ordering of EDUs. To achieve this, it suffices to order relations lexicographically, extending the definition of $<_T$ to complex units by comparing textual left bounds, and in case of equality, defining sublabels as preceding their parent CDU (which cover every case since CDU labels are non-overlapping as a consequence of the tree requirement for $\langle A, O\rangle$)[7]. Formally: $\pi <_T \pi'$ iff, letting $l_\pi, l_{\pi'}$ be the $<_T$ minimal EDU labels such that $l_\pi <^* \pi$ and $l_{\pi'} <^* \pi'$, we have either $l_\pi <_T l_{\pi'}$ or $l_\pi = l_{\pi'}$ and $\pi <^* \pi'$

A set of coherence relations instances $D \subseteq E$ is then simply ordered lexicographically ($R(\pi_1, \pi_2) < R(\pi_1', \pi_2')$ iff $\pi_1 <_T \pi_1'$ or $\pi_1 = \pi_1'$ and $\pi_2 <_T \pi_2'$). If more than one relation hold between the same arguments, the order in which they are evaluated does not have any semantic impact, so we assume an arbitrary ordering in that case.

Using the above order, we extend the definition of the interpretation f to include (CDU) labels, defining for such a label π, $f(\pi)$ denote the list of relations obtained from ordering the set $\{R(\pi', \pi'') \mid \pi'' \prec \pi\}$[8].

The language of *SDRS* formulae is the base level language extented with propositions $R(\pi, \pi')$ for $R \in \mathcal{R}, \pi, \pi' \in \Pi$. The semantic content K^S_π of a label π in S, is defined as a *SRDS* formula, as follows:

Definition 9 (Semantic content)**.** Let S be an SDRS:

- For an EDU label l, $K^S_l = f(l)$

[7] Also, $\langle A, E \cup O\rangle$ being acyclic imposes that a relation never links a label to one of his ancestors or descendants.

[8] The constraint $\pi'' \prec \pi$ purposedly focus on the second argument π'', in order to adequately capture discourse asides which must be evaluated as part of the content of π.

- For a CDU label, let $f(\pi) = \langle R^1(\pi^1_1, \pi^1_2), \ldots, R^n(\pi^n_1, \pi^n_2)\rangle$ and define
$$K^S_\pi = R^1(\pi^1_1, \pi^1_2) \wedge \cdots \wedge R^n(\pi^n_1, \pi^n_2)$$
the dynamic conjunction of the coherence relations' predicates in the scope of π.

In the following, when S is clear from context, we write K_π instead of K^S_π.

Now, it remains to define the semantics interpretation of relations. In SDRT, the full semantics of relations is left underspecified but for each relation, an expression of its main semantic consequences relative to the content assigned to its argument labels is provided. Let then $\Phi_{R(\pi,\pi')}$ denote such (underspecified) semantic constraints imposed by R over labels π and π' in S. For a veridical relation R, the following defines $[\![R(\pi,\pi')]\!]$:

$$x\ [\![R(\pi,\pi')]\!]\ x' \text{ iff } x\ ([\![\pi]\!] \circ [\![\pi']\!] \circ [\![\Phi_{R(\pi,\pi')}]\!])\ x'. \qquad \text{(Schema for veridical relations)}$$

For left-only veridical relations, the term $[\![\pi']\!]$ is deleted in the formula above, and for non-veridical relations, this simply boils done to $[\![R(\pi,\pi')]\!] = [\![\Phi_{R(\pi,\pi')}]\!]$.

From the above we see that interpretation of SDRS formulae, hence, that of the semantic content of labels, is defined in a double recursion: for a SDRS formula $\chi = R(\pi,\pi')$, $[\![\chi]\!]$ depends on $[\![\Phi_{R(\pi,\pi')}]\!]$ where $\Phi_{R(\pi,\pi')}$ is assumed to be a SDRS formula itself computed from the contents of π and π'. Now, since the contents of π and π' involve only strictly lesser label w.r.t. $\prec$, and that $\prec$ is well founded by definition, the correction of this doubly recursive definition follows.

Let us finally have a look into the semantic consequences of a few relations. We need to first precise the base level langage, its models and states: following the tradition in relations' semantics in SDRT, consider a base level language with, at least, the expressivity of the DRT fragment presented in section 1.2. We provide the semantics of the relations `Conditional`, `Elaboration`, `Alternation` and `Explanation`, to which aim it will suffice to assume a base language with a conditional operator $>$ in addition to standard dynamic cunjunction and negation, a signature Σ involving at least binary predicates for (temporal) precedence and inclusion of eventualities (resp. $\sqsubset, \subseteq \in \Sigma(2)$), and models which are tuples $\mathcal{M} = \langle D, W, *, I(k)\rangle$ with D a set of domain individuals, W a set of worlds, $* : W \times \wp(W) \mapsto \wp(W)$ a worlds selection function, and $I_k : \Sigma(k) \times D^k \times W \mapsto \{\top, \bot\}$.

The following statements are assumed to be theorems of the underlying logic and define (minimal) semantic constraints for `Conditional`, `Elaboration`, `Alternation` and `Explanation`:

Definition 10 (Some semantic consequences)**.**

$$\Phi_{\texttt{Conditional}(\pi,\pi')} \rightarrow (K_\pi > K'_\pi)$$
$$\Phi_{\texttt{Elaboration}(\pi,\pi')} \rightarrow (e_{\pi'} \subseteq e_\pi)$$
$$\Phi_{\texttt{Alternation}(\pi,\pi')} \rightarrow (K_\pi \vee K_{\pi'})$$
$$\Phi_{\texttt{Explanation}(\pi,\pi')} \rightarrow (\neg(e_\pi \sqsubset e_{\pi'}) \wedge \neg K_{\pi'} > \neg K_\pi)$$

In the above, we let e_π denote the "main" eventuality in the content of label π. The main eventuality of an EDU label l is straightforwardly defined as the single eventuality introduced by the EDU, *i.e.* the single event or state discourse referent introduced in the DRS K_l. For CDU labels, a recursive computation needs defining. We do not however fully detail this, as it requires a close investigation into event semantics taking us too far from our main concern. Let us simply assume that we can define for each type of relation how the argument's event and conditions thereon contribute to define a "sum" event, and that the main event of a CDU label is a mereological compound of the event produced by the relations it contains[9].

[9] The idea here is that, for instance, $\texttt{Alternation}(\pi_1, \pi_2)$ will produce a "sum" eventuality e on which the disjunction of the condition holding on e_{π_1} and those holding on e_{π_2} must hold, whereas a veridical relation would typically define a main eventuallity encompassing both e_{π_1} and e_{π_2} as subeventualities.

Put in less formal terms, $\texttt{Conditional}(\pi, \pi')$ simply requires that K'_π's normally follows from K_π. $\texttt{Elaboration}(\pi, \pi')$ imposes the main event of π' to be part of the main event of π. `Alternation` provides a discourse-level relational version of dynamic disjunction. Finally, `Explanation` requires two conditions to be met: first, its first argument must not temporally precede its second, then a causal relation must hold between the second (explanans) and the first argument (explanandum), here formalized as a counterfactual claim "if the explanans would not obtain, then normally neither would the explanandum".

2.3.4 The construction of SDRSs

We now know what SDRSs are and how they are semantically interpreted. The last thing that remains to describe is the process of incrementally building an SDRS from an input text. However, the precise formalization of the mechanisms involved in this process, unlike the logical forms and semantic interpretation detailed so far, is not directly a main concern of the next sections and chapters. Indeed, in subsequent sections, we will rely on SDRT's model of update and inference of relations as it stands, without advocating major theoretical additions or dwelve far into the technical details of the mechanisms at stake. For this reason, we adopt here a global and less formally detailed perspective than the one of two preceding subsections.

As previously mentioned, SDRSs' incremental construction ressorts to the glue logic. The glue logic and associated language involve two components:

- A logical language $\mathcal{L}_{glue}$ for describing (underspecified) logical forms. $\mathcal{L}_{glue}$ represents underspecification at every level, from the lexicon to the level of discourse. A class of finite models is defined, each of which includes a represensation of a (fully specified) SDRS. A (static) tarskian satisfaction relation $\models_{glue}$ formalizes what it means for a model (hence, for a fully specified structure) to satisfies constraints expressed in $\mathcal{L}_{glue}$. Moreover, $\mathcal{L}_{glue}$ has the property that every formula imposes an upper bound on the size of its minimal models (reflecting the finite number of distinct syntactic constituents stemming from the grammar. For instance, a discourse with n EDUs, imposes exactly n distinct EDU labels, and consequently no more than 2^n CDU labels in each fully specified SDRS). As a consequence satisfiability and decidability w.r.t. $\models_{glue}$ are decidable. $\mathcal{L}_{glue}$ can indeed be seen as an essentially propositional (more accurately, involving skolem normal forms without function symbols) modal[10] language.

- A non-monotonic defeasible consequence $\mathrel{|\!\sim}$, built on top of $\models_{glue}$. This defeasible consequence models the **pragmatically prefered** interpretations of a given underspecified representation. Technically, $\mathcal{L}_{glue}$ includes a modal conditional operator $>$ representing defaults, and the set of formulae φ such that $T \mathrel{|\!\sim} \varphi$ is defined using a saturation procedure consisting in, roughly, choosing a given formula a such that $T \models_{glue} a > b$ for some b, then adding a formula $a \rightarrow c$ for every c such that $T \models_{glue} a > c$, and repeating this process while the result remains a $\models_{glue}$ consistent theory. This yields a set of maximal extensions from which one can define $T \mathrel{|\!\sim} \varphi$ as equivalent to: φ is a consequence of every maximal extension. See Asher and Morreau (1991); Asher and Mao (2000) for details. $\mathrel{|\!\sim}$ can be prooved decidable as well.

The idea behind the language $\mathcal{L}_{glue}$, is to describe parts of a logical form. To this aim variables of the language are *labels* representing syntactic nodes of a logical form. $\mathcal{L}_{glue}$ can be thought of as derived from the language of SDRS formulae, introducing a countable set of variable labels which includes discourse labels, and associating every n-ary syntactic constructor c of the language of SDRS formulae[11] with an $n+1$-ary predicate P_c of $\mathcal{L}_{glue}$. Such a predicate applied to labels $l_1 \dots l_{n+1}$ expresses that the position l_{n+1}

[10] due to the defeasible modal conditional $>$.

[11] this includes the *DRS* syntactic constructor which builds a DRS from variables and conditions.

must represent a SDRS formula obtained applying the syntactic constructor c to the formulae at positions $l_1, \ldots, l_n$.

Consider as an example the discourse "*[I won't go out,]$_{\pi_a}$ [a wolf is outside.]$_{\pi_b}$*". Let S_{W} denote the SDRS for this example and assume that the content of the top label of S_{W} is $\texttt{Explanation}(\pi_1, \pi_2)$ with

$$K_{\pi_1} = \boxed{\begin{array}{|c|}\hline s \\ \hline s : \neg \Diamond \boxed{\begin{array}{|c|}\hline e \\ \hline e : \mathrm{Go\_out}(I) \\ \hline \end{array}} \\ \hline \end{array}} \quad \text{and } K_{\pi_2} = \boxed{\begin{array}{|c|}\hline x, s' \\ \hline \mathrm{Wolf(x)} \wedge s' : \mathrm{Outside}(x) \\ \mathit{Now} \subseteq s' \\ \hline \end{array}}.$$

The content of π_1 is fully described in $\mathcal{L}_{glue}$ by the following formula ($\in$ encodes the constructor used to syntactically build the DRS's universe):

$$\begin{aligned}\chi_1(\pi_1) = DRS(l^1_{Vars}, l_1, \pi_1) \wedge \in(l_{\mathrm{o}}, l^1_{Vars}) \wedge P_s(l_{\mathrm{o}}) \wedge P_{:}(l_{\mathrm{o}}, l_2, l_1) \wedge P_{\neg}(l_3, l_2) \wedge P_{\Diamond}(l_4, l_3) \\ \wedge DRS(l^2_{Vars}, l_6, l_4) \wedge \in(l_5, l^2_{Vars}) \wedge P_e(l_5) \wedge P_{:}(l_5, l_7, l_6) \wedge P_{\mathrm{go\_out}}(l_8, l_7) \wedge P_I(l_8).\end{aligned}$$

The content of π_2 is fully described by:

$$\begin{aligned}\chi_2(\pi_2) = DRS(l'_{Vars}, l'_2, \pi_2) \wedge \in(l'_{\mathrm{o}}, l'_{Vars}) \wedge P_x(l'_{\mathrm{o}}) \wedge \in(l'_1, l'_{Vars}) \wedge P_{s'}(l'_1) \wedge P_{\mathrm{Wolf}}(l'_{\mathrm{o}}, l'_2) \\ \wedge P_{:}(l'_1, l'_3, l'_2) \wedge P_{\mathrm{Outside}}(l'_{\mathrm{o}}, l'_3) \wedge P_{\subseteq}(l'_4, l'_1, l'_2) \wedge P_{\mathit{Now}}(l'_4).\end{aligned}$$

Finally, S_{W} is characterized as the unique SDRS S such that $S \vDash_{glue} \chi_1(\pi_1) \wedge \chi_2(\pi_2) \wedge P_{\texttt{Explanation}}(\pi_1, \pi_2, \pi_\top)$.

Using $\mathcal{L}_{glue}$, the inference of relations is driven by axioms of the form:

$$?(\alpha, \beta, \gamma) \wedge \texttt{conditions} > R(\alpha, \beta, \gamma).$$

The predicate ? offers a way to encode that α and β should be linked by some rhetorical relation as part of the content of γ. Here α, β and γ are free label-variables implicitely universally quantified over, whereas the π_i and l_i in the representation of the previous example must be understood as constants (in fact both result from the skolemisation of bounded variables: the first in the scope of a universal quantifier while the second in the scope of an existential one). In practice, this means that when reasoning about a given underspecified logical form χ, these axioms might be instanciated with each of the different tuples of label constants in χ.

The `conditions` can be anything expressible in the glue language. For instance, SDRT model the inference of causal relations, by introducing a predicate $\mathrm{Cause_D}$ into the *glue* language. $\mathrm{Cause_D}(\pi, \pi_1, \pi_2)$ holds if there is linguistic **evidence** in the context π (a CDU label outscoping both π_1 and π_2) that π_1 and π_2 are linked by a causal relation. Presence of such evidence, in other words, licence to infer $Cause_D$, is axiomatised using lexical information and/or world and linguistic knowledge encoded as sentences of $\mathcal{L}_{glue}$. For instance, the presence in π_1 of a predicate indicating a change of location such as *fall*, in cunjunction with a predicate describing a physical force initating motion, such as *push*, monotically entails $\mathrm{Cause_D}(\pi, \pi_1, \pi_2)$ for $\pi_1, \pi_2 \prec \pi$. $\mathrm{Cause_D}$ furnishes the basis for inferering `Explanation`:

$$?(\alpha, \beta, \gamma) \wedge \mathrm{Cause_D}(\sigma, \alpha, \beta) > \texttt{Explanation}(\alpha, \beta, \gamma). \qquad \text{((Simplified) rule for } \texttt{Explanation})$$

To handle the small example we introduced, it seems reasonable to assume part of world knowledge that situations involving a wolf outside and situations of not going out are good candidate for a causal relationship, something which can be accounted for assuming an $\mathcal{L}_{glue}$ axiom in the spirit of:

$$\Big((\mathrm{Wolf}(\nu, \mu) \wedge \mathrm{Outside}(\nu, \mu) \wedge \nu \prec \alpha) \wedge \big(P_{\neg}(\nu', \mu') \wedge P_{\mathrm{Go\_out}}(\eta, \lambda) \wedge \lambda \prec \mu' \prec \beta\big)\Big) \rightarrow \mathrm{Cause_D}(\sigma, \alpha, \beta).$$

Rules of inference need not necessary involve defaults. Coherence relations might also be lexicalized by discourse markers such as *because*. The presence of *because* in a given discourse label triggers a solid, non-cancellable inference to `Explanation` (though the exact scope of the relation might still be subject to pragmatic default reasoning).

The defeasible entailment is explicitely devised as to handle conflicting defaults by giving priority to the most specific premises[12]. This provides the modeler with a modular way to write down and refine inference rules for each relation without having to worry too much about their compatibility. For instance, SDRT advocates that `Narration` can be infered as a default relation, which, in the simplest approach, writes as:

$$?(\alpha, \beta, \gamma) > \texttt{Narration}(\alpha, \beta, \gamma). \qquad \text{(default } \texttt{Narration}\text{)}$$

The design of the default logic allows this rule to be used without further trouble in cunjunction with the rule for `Explanation` and the axiom $\texttt{Narration}(\alpha, \beta, \gamma) \rightarrow \neg\texttt{Explanation}(\alpha, \beta, \gamma')$ which transfers the temporal incompatibility of `Narration` and `Explanation` into the *glue* language: if Cause_{D} can be inferred, the more specific antecedent for inferring `Explanation` will get priority and `Narration` will be consequently blocked.

Before we conclude this section, let us briefly provide more details about the role played by the right frontier in the construction of the logical form of discourse. The right frontier constrains both anaphora resolution, and attachment sites for linking new discourse constituents to the context. These are indeed two faces of the same coin: in both cases the problem is to find a referent (be it a discourse label, a event referent, a point in time, or an individual) in the discourse universe, and in both cases available referents must be found on the right frontier. Discourse referents available for attachment are those of the right frontier, discourse referents available for anaphora resolution are those which are DRS-accessible in a label of the right-frontier. In the *glue* language, anaphoric expressions are therefore modeled as introducing a formula $l = l_1 \vee l = l_2 \vee \cdots \vee l = l_n$ where l is a new label constant representing the anaphoric expression, and the l_is are previously introduced accessible labels of the same sort as l (event, individual, dicourse labels, …).

Wrapping up all of the above, the update process is summarized below: assuming that the current context is a finite set of fully specified SDRS $\{S_1, \ldots, S_n\}$, and that $\{\chi_1, \ldots, \chi_n\}$ denote their respective description in $\mathcal{L}_{glue.}$. Let π_x be a new EDU to introduce to the discourse and $\chi_x(\pi_x)$ be the formula describing the content of π_x.

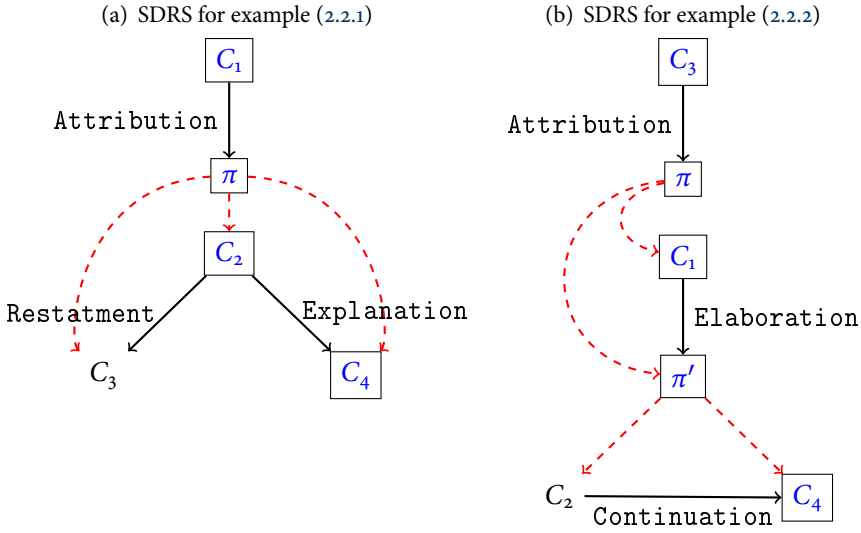
SDRT update:

- For each S_i, let avail_sites(S_i) denote the set of pairs $\langle \pi, \pi' \rangle$ such that attaching π_x to π in π' respects the right frontier constraint in S_i. Let $\kappa^x_{S_i}$ denote the formula encoding constraints on anaphora resolution: $\kappa^x_{S_i}$ imposes that each label corresponding to anaphoric expression in $\chi_x(\pi_x)$ must be equated with a label in χ_i, DRS-accessible in some discourse label on the right frontier. Define, finally $S_i{+}?(\pi, \pi_x, \pi')$ as the set of fully specified S' such that whenever $\chi_i \wedge \chi_x(\pi_x) \wedge \kappa^x_{S_i} \wedge ?(\pi, pi_x, \pi') \vdash \varphi$, we have $S' \models_{glue} \varphi$. S' is thus the (possibly empty) set of fully specified SDRSs encompassing, at once, the old information, the considered new attachment of π_x, constraints on anaphora resolution, and anything defeasibly inferable therefrom (in particular, inference of one or more coherence relations). Define finally the update of S as the set of fully-specified SDRSs obtained as the result of sequentially updating each of the S_i with each of the available attachment sites in avail_sites(S_i).

- SDRT assumes a last step in the construction of logical forms, where the set of fully-specified SDRSs obtained through updating is filtered, retaining only those which maximize *discourse coherence*. The

[12] This comes as a direct consequence of the logic validating the specifity axiom $a > c \wedge b > a \wedge b > \neg c \models_{glue} a > \neg b$

Figure 2.3: SDRSs for examples (2.2.1) and (2.2.2)



theory describes some general principles contributing to improve discourse coherence, such as the number of coherence relations inferred, the number of anaphorical link resolved, or the *quality* of infered scalar relations such as `Contrast`.

2.3.5 Examples

To conclude our introduction to SDRT, we briefly discuss the SDRSs for examples (2.2.1) and (2.2.2). The structures are displayed on fig. 2.3. We represent SDRSs with the same graphical convention as before. To stay consistent with our labeling choices in the segmentation of the examples, we keep C_i as the label of the i^{th} EDU. We use greek letters to label CDUs.

We wish to emphasize the following considerations relative to these examples:

- On fig. 2.3(a), the right scope of the `Restatement` is C_2 only, while the left scope of the `Explanation` is C_3 only. Now that we are familiar with the interpretation of SDRSs, we can appreciate how this impacts semantic interpretation, and correlatively how placing both C_2 and C_3 in the scope of the `Explanation` would differ: doing so would yield a weaker (and inaccurate) meaning representation for this part of the discourse, implying the proposition that "if Canada output of crude oils weren't shrinking, then either C_3 or C_2 or the fact that C_3 is a restatement of C_2 would normally not hold".

- In the RST Tree on fig. 2.1(a), the left subtree for the `Explanation` node, spans over C_2 and C_3. Showing that semantic scopes are, to the least, not always directly readable from an RST Tree.

- The same kind of remark can be conducted regarding the relative scope of `Elaboration` and `Attribution` in example (2.2.2).

- Example (2.2.1) is arguably ambiguous between the structure on fig. 2.3(a) and a structure where C_3 is not a subconsitutent of π (semantically, a structure where the author endorse the restatment instead of reporting it). As we have seen, this ambiguity can be accounted for in the *glue* language.

- In RST, `Attribution` is a SN-relation (the matrix clause is a satellite) whereas in SDRT it is a subordinating relation (NS-equivalent, the matrix clause dominate the reported speech). However,

Hunter et al. (2006) shows how in the framework of SDRT an evidential, veridical reading is, in many case, defeasably inferable. In such an evidential reading the matrix clause is subordinated to the reported speech, which is, intuitively what RST's `Attribution` captures.

On another note, several corpora annotated in the framework of SDRT now exist. We mention here, DISCOR (Baldridge et al., 2007), ANNODIS (Afantenos et al., 2012a) and STAC (Afantenos et al., 2012b), which have supported important empirical investigations and the main parsing efforts lead in SDRT. These 3 corpora are constituted of discurvely annotated texts from, respectively, the Wall Street Journal, the french Est Republicain, and tchat conversations between players of an online version of Settlers of Catan.

2.4 Some other approaches: Dependency graphs, Discourse DAGs, D-LTAG

We briefly discuss in this section some additional approaches to discourse structure. These approaches arguably differ from SDRT and RST insofar as they are either computationnally driven, or relative to some corpora and aim rather at capturing a certain class of structures fulfilling a given purpose than providing an abstract theoretical account of the nature of discourse structure and of what exactly it expresses. We are mostly intereseted in the constraints defining each adopted class of discourse structure, and will focus on this aspect.

We already mentioned D-LTAG in the beginning of the present chapter. D-LTAG is a grammatical formalism extending lexicalized tree adjoining grammars to discourse. Coherence relations in D-LTAG are all lexicalized by, either relaized or unrealized discourse markers. This together with the rules of composition in D-LTAG makes the set of finalized structures reducible to binary RST-Trees without nucleus-satellite annotation.

We also consider Dependency Graphs (DGs). Muller et al. (2012) derive DGs from the SDRSs of the ANNODIS corpus to get a reduced search space, simplifying automated discourse parsing. A DG is an SDRS in which there are no CDUs and there is a unique arc between any two nodes, and which is is projective[13]. Muller et al. (2012) provide a procedure ζ from SDRSs to DGs, which we slightly modify to respect the Frontier Contraint that they use. ζ works in a bottom-up fashion replacing every CDU π that is an argument of a rhetorical relation by the top-most, textually leftmost, immediate sub-constituent of π which do not appear as the right argument of any relation in π. If this violates projectivity, the CDU is instead removed by distributing relations over its internal consitutents, removing internal relations that violates projectivity. To give a simple example (with a simplified notation, the content of CDU written in square brackets): $\zeta(R([R'(a,[R''(b,c)])],d)) = \zeta(R([R'(a,b) \wedge R''(b,c)],d)) = R(a,d) \wedge R'(a,b) \wedge R''(b,c)$. Example (2.2.2) provides a more complicated example where a relation is distributed (and the `Continuation` deleted) to keep the result projective. Figure 2.4 depicts the trees for examples (2.2.1) and (2.2.2). For these examples and the vast majority of cases, output structures are Dependency Trees (DTs) *i.e.* they further verify that any node has at most one parent node (*i.e.* appears on the right of at most one relation).

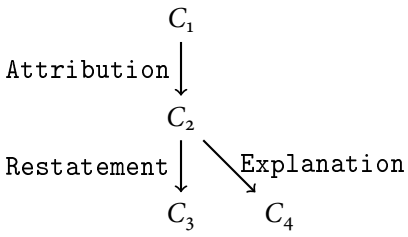
To obtain the DG depticted fig. 2.4(b), ζ is applied to the SDRS of fig. 2.3(b). By definition ζ always produces projective DGs. In that particular case, replacing π with its "head" makes $\texttt{Continuation}(C_2, C_4)$ cross $\texttt{Attribution}(C_1, C_3)$ therefore, ζ drops $\texttt{Continuation}(C_2, C_4)$ and distributes the `Elaboration` over C_2 and C_4.

Another influential contribution on the structure of discourse, and the last we mention here, is brought up by Wolf and Gibson (2005a). Wolf and Gibson propose an empirically driven and corpus-based account. They advocate generic graph structures which they use in building a corpus, the discourse Graphbank, annoted with such structures. They put forward a number of examples that they argue cannot be represented

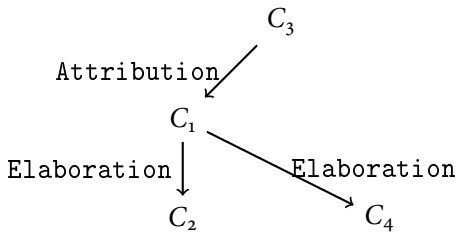
[13] *i.e.* which does not contain crossed dependencies $\cdot \overset{\frown}{\cdot\cdot\cdot\cdot} \cdot$ w.r.t. to the textual ordering of nodes

Figure 2.4: DGs for examples (2.2.1) and (2.2.2)

(a) DG for example (2.2.1)



(b) DG for example (2.2.2)



as projective tree structures, mostly due to nodes linking *via* several relations to multiple nodes, or crossed dependencies.

The set of structures used in the Graphbank corpus are very close in spirit to SDRSs, and might be identified with "relaxed" SDRSs which do not mandatorily respect the right frontier constraint (for instance, a structure like  might be encountered in Graphbank, allthough the right frontier constraint forbids such usage of a coordinating relation R_{coord}).

2.5 Questions raised

We have reviewed some of the main approaches to discourse structures, with a special focus on RST and SDRT. While each formalism proposes a perspective of its own on the nature and interpretation of coherence relations, some concerns (especially, regarding the semantic consequences of relations) are shared. We have also seen that several corpora now exist annotated with such structures: RSTTB, Discor, GraphBank[14].

The question naturally arrises then, of how exactly these annotations compare, and consequently of how corresponding theoretical views diverge and/or agree. A related problem is one of quantifying such agreement or disagreement. We have discussed differences in structural constraints governing sets of possible structures, but can we formalise all of these constraints within a common framework? Finally and more importantly, can we compare those structures and translate from one formalism to another? For instance, what kind of information should we extract from, *e.g.* an RST structure that is provided to us, assuming we are interested in an application in the framework of SDRT? Such considerations constitute the topic of the two following chapters.

[14] The Penn Discourse Treebank could also be considered as a corpus with partial dependency structures.

Chapter 3

Expressivity and comparison of discourse theories

Contents

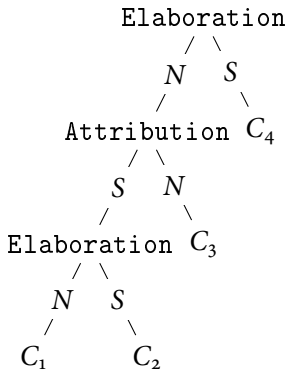
3.1 Motivation: different scopes for different interpretations

3.1.1 Differences in the scope of relations

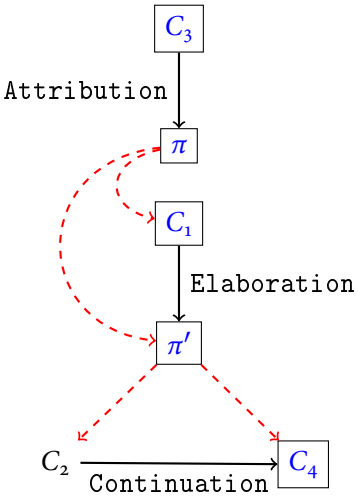
Let us start with a concrete look at inter-theoretical structural discrepancies, through a graphical summary of the treatments of one of our running examples, example (2.2.2). Recall first that we focus here on the comparison of structures, not on matching relations taxonomies. Hence the proposed methods are designed to be independant of any specific set of relations. Translations between formalisms will work as long as one applies them to structures involving relations from a single given set. Put another way, we assume that a meaningful correspondance between relations is already achieved (*e.g.* on the basis of works

Figure 3.1: Different structures for example (3.1.1)

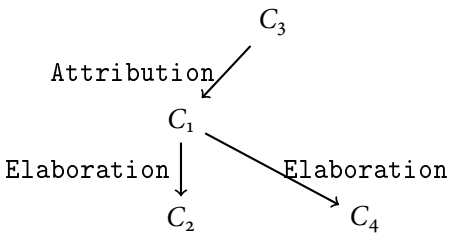
(a) RST tree for example (3.1.1) (intensional `Attribution`)



(b) SDRS for example example (3.1.1)



(c) DG for example (3.1.1)



like Sanders et al. (1992); Bateman and Rondhuis (1997); Benamara and Taboada (2015)), so that structures to compare already agree on the set of coherence relations involved. Therefore, our examples rely on a fixed set of relations involving `Explanation`, `Elaboration`, `Attribution` and the like. We also assume, for the sake of exposition, an intensional reading of `Attribution` in example (2.2.2). Consequently we consider for this example an RST tree which departs slightly from the one provided in Egg and Redeker (2010) and which is presented in fig. 2.1(b). In term of structures, this means that we use here `Attribution` as an *NS* relation, while fig. 2.1(b) implements `Attribution` as an *SN* relation, which is more classical in RST. Accordingly, we assume an SDRS with an intensional `Attribution` as well, which is already the structure provided in fig. 2.3(b). We leave it up to the interested reader to apply the necessary changes to see how, in the following, the RST tree of fig. 2.1(b) and the suitably modified version of the SDRS of fig. 2.3(b) implementing the evidential (*SN*, veridical) `Attribution` compare in our formalism. We duplicate example (2.2.2) below into example (3.1.1) and the different structures and associated formalisms are exposed on fig. 3.1.

(3.1.1)[“he was a very aggressive firefighter.]$_{C_1}$ [he loved the work he was in,"]$_{C_2}$ [said acting fire chief Lary Garcia.]$_{C_3}$. ["He couldn't be bested in terms of his willingness and his ability to do something to help you survive."]$_{C_4}$

RST provides a tree annotated with nuclearity features presented on fig. 3.1(a), which (s_1) linearly encodes. SDRT provides a different kind of structure visible on figure fig. 3.1(b), with equivalent linear encoding in (s_2). Dependency graphs (DGs) from Muller et al. (2012) give yet another representation, depicted on fig. 3.1(c), encoded by (s_3).

$$\texttt{Elaboration}_1(\texttt{Attribution}(\texttt{Elaboration}_2(C_{1N}, C_{2S})_S, C_{3N})_N, C_{4S}) \tag{s_1}$$

$$\texttt{Attribution}(C_3, \pi) \wedge \pi{:}\texttt{Elaboration}(C_1, \pi_1) \wedge \pi_1{:}\texttt{Continuation}(C_2, C_4) \tag{s_2}$$

$$\texttt{Elaboration}_1(C_1, C_2) \wedge \texttt{Attribution}(C_1, C_3) \wedge \texttt{Elaboration}_2(C_1, C_4) \tag{s_3}$$

We thus see, in the illustrative example chosen and for the relation types they agree on, that different annotation models and theoretical frameworks all represent a form of dominance, using an order relation

over discourse constituents and/or relations instances, but invoke different numbers of relation instances and assign them different arguments or different scopes, at least on the surface. This should not be much of a surprise, since the kind of dominance that a given type of structure represents need not necessary be purely semantically driven. Consequently, it is not mandatory that a structure transparently encodes the sets of arguments that relations, taken as semantic objects, must have. It is the case, by choice of design, for SDRT, but might differ for other theories. It seems unlikely however, that a salient notion of dominance between discourse constituents leaves the actual semantic scope of relations completely unrestricted. In other words, a structure of a given kind, even though it does not express a semantic representation *per se*, should at least constrain the set of compatible semantic scope for relations.

For this reason, we develop in this chapter a method of comparison based on sets of admissible scopes of relations in different types of structures. To this aim we define a central notion of interpretation into a common formalism, applicable to different types of structures, and which formally specifies the set of relations' semantic scopes admissible with respect to a given structure of a given type and its informational content. This theoretical work is important for furthering empirical research on discourse: discourse annotations are expensive. It behooves researchers to use as much data as they can, annotated in several formalisms, while pursuing prediction or evaluation in their chosen theory. We provide a theoretical basis to do this.

3.1.2 What do structures express?

Indeed, what a given structure expresses exactly is often not clear; some discourse theories are not completely formalized or lack a worked out semantics. Nevertheless, we have seen that, in all of them, rhetorical relations have important semantic consequences. Theories like SDRT or later versions of the LDM (Polanyi et al., 2004) adopt a conception of discourse structure as logical form. Discourse structures are like logical formulae and relations function like logical operators on the meaning of their arguments. Hence their exact scope has great semantic impact on a wide range of interpretative tasks in exactly the way the relative scope of quantifiers makes a great semantic difference in first order logic (these interpretative tasks involve, among others, recognizing of textual entailment, anaphora and ellipsis resolution, or the temporal, spatial interpretation of texts). By concentrating on exact meaning representations, however, the syntax-semantics interface becomes quite complex: as happens with quantifiers at the intra sentential level, discourse relations might semantically require a scope that is, at least a priori, not determined by syntactic considerations alone and violates surface order (see s_2).

Other theories like Polanyi's early version of the LDM (Polanyi, 1985; Polanyi and Scha, 1984) or D-LTAG adopt a syntactic point of view, and RST with strongly constrained (tree-shaped) structures is subject to parsing approaches (duVerle and Prendinger, 2009; Sagae, 2009; Subba and Di Eugenio, 2009) that adhere to the syntactic approach in adopting decoding strategies of syntactic parsing. In such theories, discourse structure representations, subject to syntactic constraints (*e.g.* dominance of spans of text one over another) respect surface order but do not always and unproblematically yield a semantic interpretation that fits intuitions. According to Marcu (1996), an RST tree is not by itself sufficient to generate desired predictions; he employs the *nuclearity principle*, NP, as an additional interpretation principle on scopes of relations.

We focus here on RST and SDRT, which we have introduced at length and both counts several annotated corpora. Other kind of structures such as Graphbank's graphs, or D-LTAG trees are, in terms of structures and their interpretation, very close in spirit to either RST Trees or SDRSs, and fit in our formalism as well. We will also extensively consider the DG representations of discourse.

We will first introduce in section 3.2 a language for describing semantic scopes of relations that is powerful enough to:

1. compare the expressiveness (in terms of what different scopes can be expressed) of the different formalisms considered;
2. give a formal target language that will provide comparable interpretations of the different structures at stake

Section 3.3 discusses Marcu's nuclearity principle and proposes an alternative way to interpret an RST tree as a set of different possible scopes expressed in our language. Section 3.4.1 provides intertranslability results between the different formalisms. Section 3.5 defines a measure of similarity over discourse structures of different formalisms.

3.2 Describing the scope of relations

3.2.1 Language

We provide here a language expressive and general enough to express the structures of aforementioned formalisms. All theories involve structures that one can describe using a list of rhetorical relations and their arguments, provided two parameters are set: first, the nature of the arguments, second the set of constraints that restrict the admissible lists of relations (including, *e.g.* right frontier, or requirement for tree structures). SDRT for instance, regarding the first parameter, licences complex discours units as argument of relations (*e.g.* $\pi : R_{subord}(b, c) \wedge R_{subord}(a, \pi)$), which finds a counterpart within RST Trees, where a relation may directly appear as argument of another ($R(a_N, R(b_N, c_S)_S)$) but not within dependency graphs.

To deal with the first parameter above, we remark that it suffices to list, for each instance of a discourse relation, the set of *elementary* constituents that belong to its left and right scope in order to express every kind of structure. We do this in a way that an isomorphic structure can always be recovered. Models of our common language will consists in a list of relation instances and elementary constituents, together with a set of predicates stating what is in the scope of what. As for the second point, we axiomatize each constraint in our common language, thereby describing each type of discourse structures as a theory of the language.

Formally, our scope language L_{scoping} is a fragment of that of monadic second order logic with two sorts of individuals: relation instances (i), and elementary constituants (l). Below, we assume that $\mathcal{R}$ is the set of all relation names (`Elaboration`, `Narration`, `Justification`, ...).

Definition 11 (Scoping language)**.** Let S be the two-elements set $\{i, l\}$. The set of primitive, disjoint types of L_{scoping} consists of i, l and t (type of formulae). For each of the types in S, we have a countable set of variable symbols V_i (V_l). Two additional countable sets of variable symbols $V_{\langle i,t\rangle}$ and $V_{\langle l,t\rangle}$ range over sets of individuals. These four sets of variable symbols are pairwise disjoint.

The alphabet of our language is constituted by V_i, V_s, a set of predicates, equality, connector and quantifier symbols. The set of predicate symbols is as follows:

- For each relation symbol r in $\mathcal{R}$, L_R is a unary predicate of type $\langle i, t\rangle$—we write $L_R : \langle i, t\rangle$.
- unary predicates, *sub*, *coord* and $sub^{-1} : \langle i, t\rangle$.
- binary predicates $\in_{\cdot|}$ and $\in_{|\cdot}$: $\langle l, i, t\rangle$.
- two equality relations, $=_s : \langle s, s, t\rangle$ for $s \in \{i, l\}$.

Logical connectors, and quantifiers are as usual:

- $\wedge$,$\vee$, $\rightarrow$ of type $\langle t, t, t\rangle$, $\neg$ of type $\langle t, t\rangle$

- For each type $s \in \{i, l\}$, for each variable $x \in V_s$, $\forall x{:}s$ and $\exists x{:}s$ of type $\langle t, t\rangle$.
- For each type $s \in \{i, l\}$, for each set variable symbol $X \in V_{\langle s,t\rangle}$, $\forall X{:}\langle s, t\rangle$ and $\exists X{:}\langle s, t\rangle$ are of type $\langle t, t\rangle$.

The sets Γ_τ of well-formed terms of type τ are defined, as follows:

- variables, or logical constants of type τ are in Γ_τ.
- For each symbol σ of type $\langle \tau_1, \ldots, \tau_n\rangle$ in the alphabet, for all $(t_1, \ldots, t_{n-1}) \in \Gamma_{\tau_1}\times\cdots\times\Gamma_{\tau_{n-1}}, \sigma[t_1, \ldots, t_{n-1}] \in \Gamma_{\tau_n}$.

Γ_t is the set of well formed formulae of the scope language.

The predicates $\in_{\cdot|}$ and $\in_{|\cdot}$ take a relation instance r of type i and a elementary constituent x of type l as arguments. Intuitively, they mean that x has to be included in the left (for $\in_{\cdot|}$) or right (for $\in_{|\cdot}$) scope of r. For each relation symbol R such as `Justification` or `Elaboration`, the predicate L_R takes a relation instance r has argument and states that r is an instance of the rhetorical relation R. Predicates *sub*, *coord* and sub^{-1} apply to a relation instance r, respectively specifying that r's left argument hierarchically dominate its right argument, that both are of equal hierarchical importance, or that the left one is subordinate to the right one.

Definition 12 (Scoping structure and Interpretation)**.** A *scoping structure* is an L_{scoping}-structure *i.e.* a tuple $\mathcal{M} = \langle D^i, D^l, |.|^{\mathcal{M}}\rangle$ where D^i and D^l are disjoint sets of individuals for the sorts i and l respectively. $|.|^{\mathcal{M}}$ assigns to each predicate symbol P of type $\langle \tau_1, \ldots, \tau_n, t\rangle$ a function $|P|^{\mathcal{M}} : D^{\tau_1} \times \cdots \times D^{\tau_n} \mapsto \{0, 1\}$. Variables of type $\langle i, t\rangle$ are assigned (characteristic functions of) subsets of D^i and similarly variables of type $\langle l, t\rangle$ are assigned a subsets of D^l. The predicates $=_i$ and $=_s$ are interpreted as equality over D^i and D^l respectively.

The interpretation $[\![\cdot]\!]^{\mathcal{M}}_v$ of a formula $\varphi \in \Phi_S$ is the standard (*i.e.* with "full" semantics) interpretation of a monadic second order formula w.r.t. to a model and a valuation (interpretation of first order quantifiers and connectors is as usual, quantification over sets is over all sets of individuals). Validity $\models$ also follows the standard definition.

These scoping structures offer a common framework for different discourse formalisms. We first show how to encode SDRT, RST and DGs as scoping structures, and then how to use the scoping language to axiomatise each of the 3 formalisms constituting our main interest.

3.2.2 Encoding to and decoding from scoping structures

We first proceed to encode SDRT, RST and DGs as scoping structure, and the other way around:

Fact 1. For each of the tree type of structures $T \in \{\text{SDRT}, \text{RST}, \text{DG}\}$, one can define a pair of algorithms $\langle I_T, E_T\rangle$ such that:

- From a given structure s of the theory T, I_T computes a scoping structure $s' = I_T(s)$.
- Given such a computed structure s', $E_T(s')$ retrieves the original structure s, *i.e.* $E_T(I_T(s)) = s$.
- Conversely, for each scoping structure s' such that a call to $E_T(s')$ is successful and $E_T(s') = s$ we have $I_T(E_T(s')) = s'$.

To flesh out I_T and E_T for the different theories, we need to define dominance between relation instances. Given a scoping structure $\mathcal{M}$, set L_Args$(r) = \{e \in D^l \mid (e, r) \in |\in_{.|}|^{\mathcal{M}}\}$ (the set of edus in the left scope of r); define R_Args(r) analogously (replacing $\in_{.|}$ with $\in_{|.}$). Let Args$(r) \triangleq$ L_Args(r) $\cup$ R_Args(r). Left and right dominance for relations, $\sqsubset_{.|}$ and $\sqsubset_{|.}$, are defined as follows:

$$r \sqsubset_{.|} r' \text{ iff } \text{Args}(r) \subseteq \text{L\_Args}(r')$$
$$r \sqsubset_{|.} r' \text{ iff } \text{Args}(r) \subseteq \text{R\_Args}(r')$$

Dominance $\sqsubset$ is the disjunction of the two above relations: $\sqsubset = \sqsubset_{.|} \cup \sqsubset_{|.}$. Note that these are all axiomatizable in the scoping language:

$$r \sqsubset_{.|} r' \equiv \forall z{:}l\ ((z \in_{.|} r \vee z \in_{|.} r) \rightarrow z \in_{.|} r') \text{ with } \sqsubset_{|.} \text{ axiomatized analogously.}$$
$$\text{L\_Args}(r, X) \equiv \forall z{:}l\ (z \in_{.|} r \leftrightarrow z \in X) \text{ with R\_Args}(r, X) \text{ similar}$$
$$\text{Args}(r, X) \equiv \forall z{:}l\ ((z \in_{.|} r \vee z \in_{|.} r) \leftrightarrow z \in X).$$

RST Encoding I_{RST} The NS, NN and SN relation types of RST will be respectively encoded by the predicates *sub*, *coord* and sub^{-1}. We proceed recursively. If t is an EDU e, return $M_t = \langle D_i = \varnothing, D_l = \{e\}, \epsilon\rangle$ where ϵ is the interpretation that assigns the empty set to each predicate symbol. If the root of t is a binary node instantiating a relation $R(t_{1a_1}, t_{2a_2})$, let $H_r \in \{sub, coord, sub^{-1}\}$ be the predicate that encodes the nuclearity type a_1a_2. Let $M_{t_1} = \langle D_i^1, D_l^1, |.|^1\rangle$ and $M_{t_2} = \langle D_i^2, D_l^2, |.|^2\rangle$ be the scoping structures returned by the alogithm for t_1 and t_2 respectively. The algorithm returns $M_t = \langle D_i^1 \cup D_i^2 \cup \{r\}, D_l^1 \cup D_l^2, |.|^{M_t}\rangle$ where r is a 'fresh' relation instance variable not in D_i^1 or D_i^2, and $|.|^{M_t}$ is updated in the appropriate fashion to reflect the left and right arguments of r and ensure that H_r and L_R are the one and only type and label predicate that r satisfies. Finally, if the root of t is an acyclic graph $\langle V, E\rangle$ where E has more than one relation (RST schema 3 or 4), for each $R_i(t_i^1, t_i^2) \in E$, retrieve the N/S annotation a_1, a_2 for t_i^1 and t_i^2 in V, then proceed as above to recursively compute the scoping-structures M_i encoding the relation (taking care to introduce a 'fresh' relation instance r_i into the model for each relation in E). Finally return the merge of the models M_i.

RST Decoding E_{RST} Given a finite scoping structure $\mathcal{M} = \langle D^i, D^l, |.|^{\mathcal{M}}\rangle$, for each relation instance r compute the left arguments of r and its right arguments. Then identify Label(r), the unique relation symbol R such that $r \in |L_R|^{\mathcal{M}}$. If that fails, the algorithm fails. Similarly retrieve the right nuclearity schema from the adequate predicate that applies to r. If $D^i = \varnothing$ and the structure is a single EDU e, returns e. Otherwise compute the dominance relations for r. If $\mathcal{M} = I_{\text{RST}}(t)$ for some RST Tree t then there is at least one maximal relation instance for the dominance relation. If t the root node of t is a binary node (schema 1 or 2), there is exactly one maximal element in the dominance relation. If there is none, then we return fail.

1. If there is exactly one, call it r_{top}, recursively compute the two RST Trees t_1 and t_2 for the argument spans of r_{top} by applying the alogrithm to the restriction of $\mathcal{M}$ to the set of relation instances in D^i that are, respectively, left- and right-dominated by r_{top}. Then use the spans of t_1, t_2 and the nuclearity information and label for r_{top} to reconstruct the root node from the two recursively computed subtrees and return the result.

2. If there are k maximal relations, the root node of the encoded RST Tree must result from a schema 3 or 4 and be a directed acyclic graph $V = \langle t_{0a_0}, \ldots, t_{na_n}\rangle, E = \{R_i(t_{k_i}, t_{l_i}) \mid 0 \leq i \leq k\}$. Each relation instance r_i maximal for dominance must encode one of the relation in E, so restrict D^i to r_i and the relations it dominates, then proceed exactly as in 1. to retrieve r_i's label (R_i), nuclearity, argument

spans and their substructures. Once this is done, if no failure has been encountered, reconstruct, if the retrieved cardinality, nuclearity and argument spans of the relations matches, the schema application for the root node and returns it. Otherwise, the root node does not result from the application of a valid RST schema and the algorithm fails.

SDRT Encoding and Decoding, I_{SDRT}, E_{SDRT}: Encoding and decoding for SDRT is quite similar to the case of RST, so we sketch it with less details: the idea is once again to rely on the recursivity of the input structure.

The encoding is rather simple, it amounts to list for each relation in the input SDRS the set of elementary units that it scopes over, then encode this and the relation's label and subordinating/coordinating type into the interpreation of the adequate predicates. For instance, we can proceed top down as we dit for RST: A SDRS s can be seen as a complex constituent π_{top} that contents a graph $g = \langle V, E\rangle$ whose edges are relations holding between sub-constituents, simple or complex as well. First come up with an encoding of the set E of all edges that hold between two sub-constituents of s, *i.e.* a structure $\mathcal{M} = \langle D_i = E_i, D_l = V, |\cdot|^{\mathcal{M}}\rangle$ where the following holds: for each edge $e \in E_i$, $e \in |L_R|^{\mathcal{M}}$ and $e \notin |L_{R'}|^{\mathcal{M}}$ for $R \neq R'$, this encodes the label of e. Similarly one and only one predicate among *sub* and *coord* holds of e. Finally, $|\in_{\cdot|}|^{\mathcal{M}}$ and $|\in_{|\cdot}|^{\mathcal{M}}$ consists of all the pairs (x, e) of left and right nodes x of the edges $e \in E$. Then, for each complex immediate sub-constituent of s in D_l, update $\mathcal{M}$ as follows: for c such a subconsituent, recursively compute its encoding M_c, then add everything of M_c to M, finally remove c from M but add instead for each relation r scoping over c to the right (left), all the pairs $\{(x, r) \mid x \text{ is a constituent in } M^c\}$.

The crucial point behind the decoding is, as in the RST case, to retrieve the hierarchical structure using only the set of elementary units over which relations scope, which is what the scoping structure informs us about. Again, computing dominance between relations does the trick. The relations which are maximal for dominance must form the content of the top-label, which can thus be retrieved. The recursive steps (*i.e.* successively moving to inner constituents) are achieved by restricting the domain of relation instances D^i to the left or right dominated relations of each of the maximal relations.

DG: Dependency graphs are syntactically a special case of SDRSs; there is only one CDU whose domain is only EDUs.

So far we have not really used the scoping language, only scoping structures as a pivot formalism. We can for instance imagin encoding an SDRS into a scoping structure and decode it into an RST Tree. Of course this might fail if the input SDRS fails to meet the structural requirements of RST Trees. As a consequence our decoding algorithms implements partial functions. Indeed only those structures obtainable from encoding an RST tree will yield an RST Tree when fed to the RST decoding algorithm. One can then wonder what this set of structures exactly is. The next subsection provides an answer to this question, which also constitutes our first use of the scoping language.

3.2.3 Structural constraints axiomatized

The scoping language allows us to axiomatize the three classes of scoping structures corresponding to RST Trees, SDRSs and DGs. As not all scoping structures obey these axioms, our language is strictly more expressive than any of these discourse formalisms.

As a first example of an axiom, the following formula expresses that a relation cannot have both left and right scope over the same elementary constituent:

Strong Irreflexivity:

$$\forall r{:}i \forall x{:}l\ \neg(x \in_{\cdot|} r \wedge x \in_{|\cdot} r) \qquad (A_0)$$

Strong irreflexivity entails irreflexivity; a given relation instance cannot have the same (complete) left and right scopes. All discourse theories validate A_0.

Below, we define axioms (A1-A9) that axiomatize the structures corresponding to RST, SDRT and DGs. Axiom A_1 says that every discourse unit is linked via some discourse relation instance. Axiom A_2 insures that all our relation instances have the right number of arguments; Axioms A_3 and A_4 ensure acyclicity and no crossing dependencies. A_5a and A_5b restrict structures to a tree-like dominance relation with a maximal dominating element, while A_6 defines the Right Frontier constraint for SDRT, and A_7 fixes the domain for SDRT constraints on CDUs. A_8 ensures that no coordinating and subordinating relations have the same left and right arguments, while A_9 provide the restrictions needed to define the set of DGs:

We need to write down some elementary definitions in the scoping language before we axiomatize the different constraints: we first introduce $scope(r, x)$ as a shortcut for $x \in_{\cdot|} r \vee x \in_{|\cdot} r$. We assume that we have access to the textual order of EDUs as a strict linear ordering $<_t : \langle l, l, t\rangle$ over EDUs. We also appeal to the notion of a chain of relations instances forming a structure $x_1 \rightarrow^{r_1} x_2 \rightarrow^{r_2} \ldots \rightarrow^{r_n} x_n$ with arguments $x_1 \ldots x_n$ constituted only of EDUs in X. We write this as $Chain(R, X)$ for a set of relations R and a set of EDUs X. $Chain(R, X)$ can be expressed in MSO and we provide below a possible expression. The first line states that every relations in R scopes over a subset of X, the second that, for all but a single relation $r \in R$ (the leftmost one), any relation $r' \in R$ must have a left scope which equals[15] the right scope of another relation of R. We do not impose irreflexivity nor acyclicity, as it the purpose of A_0 and A_3 respectively.

$$(\forall r{:}i, x{:}l\ (r \in R \wedge scope(r, x)) \rightarrow x \in X) \wedge$$
$$\exists r{:}i\ \Big(r \in R \wedge \forall r'{:}i\ (r' \in R \wedge r' \neq r) \rightarrow \exists r''{:}i\ \Big(r'' \in R \wedge \forall X'{:}\langle l, t\rangle\ \mathrm{R\_Args}(r', X) \leftrightarrow \mathrm{L\_Args}(r'', X)\Big)\Big)$$

To handle RST relations with multiple satellites, we define a *nest*: $Nest(X, R)$ iff all $r \in R$ have the same left argument in X but take different right arguments in X. Finally, we define CDUs, conforming the definition of SDRS, we define a predicate EDU testing whether a set of constituent is as singleton and a predicate CDU testing whether a set of elementary constituents is fed as an argument to some relation, and closed for incoming relations:

$$\mathrm{EDU}(X) \equiv \exists! x : l\ x \in X$$
$$\mathrm{CDU}(X, R) \equiv \exists r{:}i\ (\mathrm{L\_Args}(r, X) \vee \mathrm{R\_Args}(r, X)) \wedge \forall r'{:}i\ ((\forall x{:}l\ (x \in_{|\cdot} r' \rightarrow x \in X)) \rightarrow r' \in R)$$

The definition for CDU does not include the top label, but it is dispensable to the expression of the structural axioms which follow.

Axiomatization

$$\forall x{:}l \exists r{:}i\ (x \in_{\cdot|} r \vee x \in_{|\cdot} r) \qquad (A_1\text{:Weak Connectedness})$$
$$\forall r{:}i \exists x, y{:}l\ (x \in_{|\cdot} r \wedge y \in_{\cdot|} r) \qquad (A_2\text{ :Properness of the relation})$$
$$\forall X{:}\langle l, t\rangle\ (X \neq \varnothing \rightarrow \exists y{:}l\ y \in X \wedge \forall r{:}i\ ((\forall z : l\ scope(r, z) \rightarrow z \in X) \rightarrow \neg y \in_{\cdot|} r))$$
$$(A_3\text{ :Acyclicity or Well Foundedness})$$

No crossing dependencies using the textual order $<_t$ of EDUs:

$$\forall x, y, z, w{:}l\ (x <_t y <_t z <_t w) \rightarrow \forall r, r'{:}i\ \Big((x \in_{\cdot|} r \wedge z \in_{|\cdot} r \wedge y \in_{\cdot|} r' \wedge w \in_{|\cdot} r') \rightarrow (r' \sqsubset_{|\cdot} r \vee r \sqsubset_{\cdot|} r')\Big). \qquad (A_4)$$

[15] The "equality" relation on sets used here can be expressed in the language as reciprocal inclusion. ZFC axiom of extensionality ensures then that this relation is indeed interpreted as equality on sets of individuals.

RST tree structures:
RST "treeness" is a little tricky to axiomatize (again due to schema 3 and 4), we will follow the same lead that lies behind E_{RST}: the idea is that the set of relations maximal for dominance gives us the internal structure of the "root" node of the tree, and must therefore instantiate a schema. Each of these relations r introduces two children nodes (one on the left and one on the right), the internal structure of which, again, is provided by the set of relations maximal within all relations left, or right dominated by r, according to the position of the considered children node w.r.t. r. These relations must in turn instanciate a schema, and so forth moving to lower-level nodes. To express this, we use quantification over sets to retrieve all "nodes" and internal topmost relations in the previous sense, *i.e.* either the set of relations maximal for dominance, or sets of relations maximal for dominance **relative to left or right dominance by some given parent relation**. Then we impose that these set of relations instanciate schemas. We use the following predicates:

$$\begin{aligned} node(X,R) \equiv & (EDU(X) \wedge R = \varnothing) \vee (\forall x : l\ x \in X \wedge \forall r : i\ r \in R \leftrightarrow \neg\exists r' : i\ r \sqsubset r') \\ & \vee ((\forall x : l, r : i\ (r \in R \wedge scope(r,x)) \rightarrow x \in X) \wedge \\ & \bigvee_{dominated \in \{\sqsubset_{\cdot|}, \sqsubset_{|\cdot}\}} \exists r_o : i\ (\forall r : i\ r \in R \leftrightarrow r\ dominated\ r_o \wedge \neg\exists r' : i\ (r' \sqsubset r_o \wedge r \sqsubset r'))) \end{aligned}$$

and two predicates $Schema_3(X,R)$ and $Schema_4(X,R)$ telling us whether the set of relations instances R results from the application of a RST schema of type 3 or 4 to X. These predicates can be axiomatised in MSO, however their exact content depends on the set of schemas of type 3 and 4 which we have not fully specified. We thus describe the general form this predicates might take:

$$Schema_3(X,R) \equiv Chain(X,R) \wedge \forall r : i\ (r \in R \rightarrow coord(r)) \wedge \bigvee_{R \in \{\texttt{Sequence}, \ \ldots\}} \forall r : i\ (r \in R \rightarrow L_R(r))$$

$$\begin{aligned} Schema_4(X,R) \equiv Nest(X,R) \wedge (\exists r_1 : i, r_2 : i\ (\forall r : i\ & (r \in R \rightarrow (r = r_1 \vee r = r_2)) \\ & \wedge sub^{-1}(r_1) \wedge sub(r_2) \wedge L_{\texttt{Motivation}}(r_1) \wedge L_{\texttt{Enablement}}(r_2))) \\ & \vee \ldots \end{aligned}$$

The condition ensuring RST tree-like structures are below[16]:

$$\forall R : \langle i,t\rangle, X : \langle l,t\rangle\ node(X,R) \rightarrow (EDU(X) \vee (\exists! r : i\ r \in R) \vee Schema_3(X,R) \vee Schema_4(X,R)) \quad (A_5a)$$

Binary RST Trees are axiomatized by taking the same axiom and assuming further that $Schema_{3,4} \equiv \bot$. In this case, the axioms simply says that

a) there is a unique maximal relation (the root node) and

b) every relation has a unique maximal left descendant and a unique maximal right descendant.

Moreover, assuming strong irreflexivity (A_0) and well foundedness (A_3) (or finite structures), this entails existence of a unique parent relation for dominance[17]:

$$\begin{aligned} \forall X : \langle l,t\rangle\ ((EDU(X) \rightarrow \exists! r : i\ (\mathrm{L\_Args(r,X)} \vee \mathrm{R\_Args(r,X)})) \wedge & \qquad (3.1) \\ \forall r : i\ (\mathrm{Args}(r,X) \rightarrow \exists! r' : i\ (\mathrm{L\_Args}(r',X) \vee \mathrm{R\_Args}(r',X)))). & \qquad (A_5b) \end{aligned}$$

[16] Notice that existence of a root node spanning over every elementary edus follows from A_1 (weak connectedness)

[17] Assume no type 3 or 4 schemas and a node r_o with two ancestors incomparable for dominance, r and r'. Using A_5a, we can show that their must be a relation s_o dominating both r and r'. Either there is a relation $s_\perp \sqsubset s_o$ which "separates" r from r', *i.e.* $r \sqsubset_{\cdot|} s_\perp$ and $r' \sqsubset_{|\cdot} s_\perp$ (or the other way around), or r and r' must both be maximal for dominance relative to the scope of some relation dominated by s_o. In the latter case, using A_5a again, we get another majorant s_1 between r-r' and s_o. Iterating the process show that there must exist a separating relation $s_\perp$: otherwise we could construct an infinite decreasing sequence $s_o, s_1, \ldots$ of distinct relations. But then $s_\perp$ violates strong irreflexivity since by definition r and r' both scope over the arguments of r_o.

Hence we recover binary RST Trees.

Right Frontier:

recall that the right frontier constraint formally expresses that any DU C (elementary or complex) attaches only to elements of the right frontier of the sub-SDRS formed by elements preceding C in textual order. In order to write this in the scoping language, we need to express the restriction to textually preceding elements: let $C:\langle l,t\rangle$ denote a set of EDUs, we overload $<_t$ with the type $\langle l,\langle l,t\rangle,t\rangle$ defining $x <_t C$ as "x textually precedes every units of C". We then define accessibily for attachment in the SDRS restricted to constituents in X. Let $\text{last}(x,X)$ hold iff x is the last element of X in textual order.

$$\begin{aligned} Acc(C,X) \equiv \exists a{:}l\ (a \in C \wedge \exists R{:}\langle i,t\rangle\ (&Chain(R,X) \wedge \forall r{:}i\ (r \in R \rightarrow sub(r)) \\ &\wedge \exists r'{:}i\ r' \in R \wedge scope(r',a) \\ &\wedge \exists x{:}l, r''{:}i\ (r'' \in R \wedge scope(r,x) \wedge \text{last}(x)))) \end{aligned}$$

this can be glossed as "C is accessible in X iff C contains an EDU a such that there is a chain of subordinating relations in X linking a and the last EDU of X in textual order. From there, the right frontier constraint simply asks that every DU or EDU links to an accessible constituant in the approriately restricted SDRS. To quantify over DUs here, it suffices to quantify over subsets of the set of EDUs which constitute the right scope of some relation:

$$\begin{aligned} \forall C{:}\langle l,t\rangle \forall r:i\ (&\text{R\_Args}(r,C) \\ &\rightarrow (\forall X{:}\langle l,t\rangle, L{:}\langle l,t\rangle\ ((\forall x:l\ x <_t C \leftrightarrow x \in X) \wedge \text{L\_Args}(r,L)) \\ &\qquad \rightarrow Acc(L,X))) \end{aligned} \tag{3.2}$$

(A_6: Right Frontier Constraint)

CDUs or EDUs and no overlapping CDUs:

$$\begin{aligned} &\forall r:i, X{:}\langle l,t\rangle\ (X = \{\text{L},\text{R}\}\text{\_Args}(r) \rightarrow (EDU(X) \vee \exists R{:}\langle i,t\rangle\ CDU(X,R))) \wedge \\ &\forall X{:}\langle l,t\rangle, Y{:}\langle l,t\rangle, R{:}\langle i,t\rangle, R'{:}\langle i,t\rangle\ (CDU(X,R) \wedge CDU(Y,R') \rightarrow (R \cap R' \neq \varnothing \rightarrow (R \subseteq R' \vee R' \subseteq R))) \end{aligned}$$

(A_7)

The same arguments cannot be linked by subordinating and coordinating relations. The formal axiom (A_8) is trivial.

DGs:

Finally, two axioms for restricting SDRSs to dependency graphs:

$$\forall r{:}i, x{:}l, y{:}l\ ((x \in_{.|} r) \wedge y \in_{.|} r)) \vee (x \in_{|.} r) \wedge y \in_{|.} r))) \rightarrow x = y \qquad (A_9a\text{:No CDUs.})$$

$$\forall r{:}i, r'{:}i, X{:}\langle l,t\rangle, Y{:}\langle l,t\rangle\ (\text{L\_Args}(r,X) \wedge \text{R\_Args}(r,Y) \wedge \text{L\_Args}(r',X) \wedge \text{R\_Args}(r',Y)) \rightarrow r = r'$$

(A_9b :unique arc)

We note that as a consequence of A_4, A_5a and A_5b we have no discourse aside or non-contiguous span:

$$\begin{aligned} \forall x:l, y:l, r:i\ &(x \in_{.|} r \wedge y \in_{.|} r \wedge x \neq y) \\ &\rightarrow \neg\exists r':i, z:l\ (x \in_{.|} r' \wedge z \in_{|.} r' \wedge \neg(z \in_{.|} r \vee z \in_{|.} r)) \end{aligned}$$

We also note that A_5a and A_5b entail A_7, A_8 and A_9b, though not vice-versa.

Using the above axioms, one can show:

Fact 2.

1. The theory $T_{(bin-)RST}$=$\{A_0, A_1, A_2, A_3, A_4, A_5 a, (A_5 b), A_8\}$ characterizes (binary) RST trees in the sense that:
   - E_{RST} applied to any structure M such that $M \vDash T_{(bin-)RST}$ yield an (binary) RST Tree.
   - for any RST Tree t, $I_{RST}(t) \vDash T_{RST}$.
2. The theory T_{SDRT}=$\{A_0, A_1, A_2, A_3, A_6, A_7, A_8\}$ similarly characterizes SDRSs.
3. The theory T_{DG}=$T_{SDRT} \cup \{A_9 a, A_9 b\}$ similarly characterizes DGs.

3.3 Immediate vs. mediated intepretation

The previous section defined the set of scope structures as well as the means to import, and then retrieve, RST trees, DGs, or SDRSs into, and from, this set. Some of these scope structures export both into RST and SDRT, yielding a 1 to 1 correspondence between a subset of SDRT and RST structures. Moreover, this correspondence is "meaningful": intuitively, it captures more than just any computable bijection between the two countable set of structures would. What, then, does it precisely capture? In mathematics, an isomorphism is a special kind of bijection that *preserves* structural operations or properties; as we shall see, our correspondence preserves the *immediate interpretation* of relations' scope.

3.3.1 Immediate Interpretation

Consider a scope structure $\mathcal{M}$ (validating A_0, A_1, A_2). Whether a relation instance r, labeled by relation name R holds between two discourse units in $\mathcal{M}$ depends on the semantic content of its left and right arguments. These are recursively described by L_Args(r), all relations r' such that $r' \sqsubset_{\cdot|} r$, R_Args(r) and all relations r' such that $r' \sqsubset_{|\cdot} r$. Algorithm I_T computes what we call the *immediate* interpretation of an input structure. Intuitively, in this interpretation the semantic scope of relations is directly read from the structure itself; a node $R(t_1, t_2)$ in a RST Tree expresses that R holds between contents expressed by the whole substructures t_1 and t_2. Similarly, for SDRT and DGs, the immediate interpretation of an edge $\pi_1 \rightarrow_R \pi_2$ is that R holds between the whole content of π_1 and π_2.

Definition 13 (Immediate interpration)**.** For an SDRS (resp. DG, RST Tree) s, the immediate interpretation of s is the scoping structure $[\![s]\!]^i = I_{SDRT(\text{resp. } DG,RST)}(s)$. In the case of SDRT, this interpretation is standard and we simply write $[\![s]\!] = [\![s]\!]^i$.

While this immediate interpretation is standard in SDRT, it is not in RST. Consider again example (3.1.1) from the introduction or:

(3.3.1) [In 1988, Kidder eked out a $ 46 million profit,]$_{C_{31}}$ [mainly because of severe cost cutting.]$_{C_{32}}$ [Its 1,400-member brokerage operation reported an estimated $ 5 million loss last year,]$_{C_{33}}$ [although Kidder expects to turn a profit this year]$_{C_{34}}$ (RST Treebank, wsj_0604).

(3.3.2) [Suzanne Sequin passed away Saturday at the communal hospital of Bar-le-Duc,]$_{C_3}$ [where she had been admitted a month ago.]$_{C_4}$ [...] [Her funeral will be held today at 10h30 at the church of Saint-Etienne of Bar-le-Duc.]$_{C_5}$ (Annodis corpus, translated).

These examples involve what are called *long distance attachments*. Example (3.3.1) involves a relation of `Constrast`, or `Comparison` between C_{31} and C_{33} but which does not involve the contribution of C_{32} (the costs cutting of 1988). Example (3.3.2) displays something comparable. A causal relation like `Result`, or

at least a temporal `Narration` holds between C_3 and C_5, but it should not scope over C_4 if one does not wish to make Sequin's admission to the hospital a month ago a consequence of her death last Saturday. Finally in example (2.2.2) C_4 elaborates on C_1, but not on the fact that C_1 is attributed to chief Garcia, so the corresponding elaboration relation should not scope over C_3.

It is impossible however, to account for long distance attachments using the immediate interpretation of RST trees. Example (3.3.1), for instance, also involves an explanation relation between C_{31} and C_{32}, which should include none of C_{33} or C_{34} in its scope. Since C_{31} is in the scope of both the explanation and the contrast relation, Axiom $A_5 a$ of the previous subsection entails than an RST tree involving the two relations has to make one of the two relations dominates the other.

3.3.2 Nuclearity Principle(s)

Marcu's Nuclearity Principle (NP, Marcu, 1996) provides an alternative to the immediate interpretation and captures some long distance attachments (Danlos, 2008; Egg and Redeker, 2010). According to the NP, a relation between two spans of text, expressed at a node of a RST Tree should hold between the most salient parts of these spans. *Most salient part* is recursively defined: the most salient part of an elementary constituent is itself, for a multinuclear schema $R(t_{1N}, t_{2N}) \dots R(t_{k-1N}, t_{kN})$ its most salient part is the union of the most salient parts of the t_i[18]. Following Egg and Redeker (2010), the NP, or *weak* NP is a constraint on which RST trees may correctly characterize an input text; it is not a mechanism for computing scopes. Given their analysis of example (2.2.2) reported in the introduction, NP entails that $\texttt{Elaboration}_1$ holds between C_1 and C_4, accounting for the long distance attachment, and that `Attribution` holds between C_1 and C_4 which meets intuition in this case. There is however no requirement that `Attribution` do *not* hold between the wider span $[C_1, C_2]$ and C_3, as there is no requirement that $\texttt{Elaboration}_1$ does not hold between $[C_1, C_2, C_3]$ and C_4. In order to accurately account for example (2.2.2), the former must be true and the latter false.

However, this interpretation of NP together with an RST tree does not determine the semantic scope of all relations. Danlos (2008) reformulates NP as a *Mixed* Nuclearity Principle (MNP) that outputs determinate scopes for a given structure. The MNP requires for a given node, that the most salient parts of his daughters furnish the **exact** semantic scope for the relation at that node. The MNP transforms an RST tree t into a scope structure $\mathcal{M}_t$, which validates $A_0 - A_3$ but also A_6.[19], A_7 and A_8. Hence $\mathcal{M}$ could be exported back to SDRT and the MNP would yield a translation from RST-trees to SDRSs.

But when applied to the RST Treebank, the MNP yields wrong, or at least incomplete, semantic scopes for intuitively correct RST Trees. The mixed principle applied to the tree of (s_1) gives the `Attribution` scope over C_1 only, but not C_2, which is incorrect. Focusing on the attribution relation which is the second most frequent in the RST Treebank, we find out that, regardless of whether we assign `Attribution`'s arguments S and N or N and S, this principle makes wrong predictions 86% of the time in a random sampling of 50 cases in which we have attributions with multi-clause second argument spans. Consider example (2.2.1) again.

Applied to the annotated RST Tree for this example (fig. 2.1(a)), the MNP yields an incorrect scope of the attribution relation over C_2 only, regardless of whether the attribution is annotated N-S or S-N.

The idea behind the weak NP provides a better fit with intuitions. The principle gives *minimal* semantic requirements for scoping relations; everything beyond those requirements is left *underspecified*. We formalize this as the *relaxed* Nuclearity Principle (RNP), which does not compute one structure where each relation is given its exact scope, but **a set** of such structures.

[18] Except for `Sequence` which only retains the most salient part of t_k

[19] That A_6 is valid in the resulting model is not immediate. Assume a multinuclear (coordinating) relation instance r has scope over x_n and x_{n+k} later in the textual order. Then it is impossible to attach with r' a later found constituent x_{n+k+l} to x_n alone, for it would require that x_{n+1} escapes the scope of r' from the MNP which it will not do by multinuclearity of r.

The target structures are not trees any more, but we want them to still reflect the hierarchical information present in the RST Tree. We therefore define a notion of *weak dominance* over structures of the scoping language: for two sets of constituents, Y weakly dominate X, written $X \trianglelefteq Y$, iff $X \subseteq Y$ or there is a relation subordinating X to Y. Weak dominance is given by transitive closure $\trianglelefteq^*$ of $\trianglelefteq$. For two relations, $r' \trianglelefteq_{\cdot|} r$ iff the left argument of r weakly dominates both arguments of r'. $\trianglelefteq_{|\cdot}$ is symmetrically defined. Recall now that the immediate interpretation of an RST Tree t simply consists in reading semantic scopes directly from the tree-descendance in t, *i.e.* transforming left or right descendance in t into left or right-dominance $\sqsubset_{\cdot|}$, $\sqsubset_{|\cdot}$ in a scoping structure. The idea behind the RNP can be formulated analogously, but this time, left and right tree descendance are respectively interpreted into left and right **weak** dominance in a scoping structure. Unlike the strong dominance constraints stemming from the immediate interpretation, the weak dominance ones are generaly realized in more than a single scoping structure. Still, the set of all structures s such that there exists an RST tree t admitting s as a member of its RNP interpretation can be axiomatized as well: one has to replace $A_5 a$ with a weakened version. To this aim the predicate *node* used in $A_5 a$'s formulation must be adapted to use weak dominance instead of strong dominance, and every node in that sense must either be an EDU, have a maximal relation for **weak** dominance, or conform to the weakened version of schema 3 or 4 (where *Chain*s and *Nest*s are tested only relative to relations nuclei). We let A_5^W denote this weakened version of $A5a$. The RNP interpretations maps any RST Tree t to a set of structure verifying $A_0, A_1, A_2, A_3, A_4, A_5^W$. Below, we formally define this RNP mapping:

Definition 14 (Relaxed Nuclearity Principle)**.** We assign to an RST Tree t a formula of the scoping language $\varphi_t = \exists \bar{x} \exists \bar{r}\ \psi_t \cup \Gamma_t$ such that:

1. $\bar{x}$ consists is a set of variable of type l, one for each EDU in t. $\bar{r}$ consists is a set of variable of type i, one for each relation instance in t.

2. ψ_t is a formula specifying that all individuals quantified in $\bar{x}$ and $\bar{r}$ are pairwise distinct, and that there is no other individuals that the ones just mentioned. ψ_t also specifies for each node n of t, the relations $R_n^1(t_1^o, t_2^o) \dots R_n^k(t_1^k, t_n^k)$ constituting of the substructure of n are labeled with the adequate relation symbols $R^o \dots R^k$ and relation type (subordinating if $NS \dots$).

3. Γ_t encodes the nuclearity principle applied to t: for every relation $R(s_1, s_2)$ appearing in the substructure of some node s of t, Γ_t specifies that the corresponding relation instance variable r must scope over EDUs in s only, and furthermore must left (resp. right) weakly dominate every relation in the substructure of s_1 (resp. s_2). Note that this entails that r includes it its left (resp. right) scope the nucleus(nuclei) of s_i[20].

The interpretation $[\![t]\!]^{\text{RNP}}$ is defined as the set of structures $\mathcal{M}$ that validate φ_t and $A_0, A_1, A_2, A_3, A_4, A_5^W$ (they all have $|t|$ individuals, as fixed by ψ_t). Moreover, it can be shown that each model of this set validates T_{SDRT}; so we have a interpretation of an RS-Tree into a set of SDRSs.

3.3.3 Illustrative example

Consider example example (3.1.1) again, with the RST Tree of fig. 3.1(a). The table below lists the scoping structures for the different interpretations. To keep things readable, we simply represent the scope scoping structure as a list of relations instances in the format $R([args_1, args_2])$, indicating a relation instance r labeled with R and scoping on the left (resp. right) over the set of EDUs arg_1 (resp.) $args_2$. We also factorise some structures using the notation $[args] \mid [args']$ to denote several distinct structures at once: one with

[20] This is obtained through an easy induction.

args as argument to the targeted relation, the other that has arg'. We drop the subordinating/coordinating nature of relations which is the same for each of the structures listed below and fig. 3.1(b).

We first display the immediate and Mixed NP interpretations:

Immediate Interpretation:	MNP:
$\texttt{Elaboration}([C_1],[C_2])$ $\texttt{Attribution}([C_1,C_2],[C_3])$ $\texttt{Elaboration}([C_1\text{–}C_3],[C_4])$	$\texttt{Elaboration}([C_1],[C_2])$ $\texttt{Attribution}([C_1],[C_3])$ $\texttt{Elaboration}([C_3],[C_4])$

The RNP interpretations are listed below:

RNP:	
$\texttt{Elaboration}([C_1],[C_2])$	$\texttt{Elaboration}([C_1],[C_2])$
$\texttt{Attribution}([C_1],[C_3])$	$\texttt{Attribution}([C_1,C_2],[C_3])$
$\texttt{Elaboration}([C_3] \mid [C_1,C_3] \mid [C_1\text{–}C_3],[C_4])$	$\texttt{Elaboration}([C_3] \mid [C_1\text{–}C_3],[C_4])$

This example and associate intepretations illustrate some noticeable facts:

- Both the immediate and the MNP interpratation are admissible under the RNP.
- The following interpretation, with overlapping consituents

$$\begin{aligned}&\texttt{Elaboration}([C_1],[C_2]\\&\texttt{Attribution}([C_1,C_2],[C_3])\\&\texttt{Elaboration}([C_1,C_3],[C_4])\end{aligned}$$

  is **not** admissible under the RNP: $[C_1,C_2]$ is not weakly dominated by $[C_1,C_3]$. As a consequence $\texttt{Elaboration}([C_1,C_3],[C_4])$ does not weakly dominate $\texttt{Attribution}([C_1,C_2],[C_3])$ on the left as the RNP requires. More generally, the RNP forbids overlapping constituents. In fact it is sufficient to ensure that all target structures comply to A_6, A_7, A_8 and $A_9 b$.

Now that we have formally defined the different interpretations of an RST Tree, we turn in the next section to the interpretation of DGs and show how it can be related to the RNP on RST Trees.

3.4 Relation between RST Trees and DGs

3.4.1 Interpretation of DGs

DGs are a restriction of SDRSs to structures without complex constituents. So the ζ function of subsection 2 can transform distinct SDRSs transform into the same DG with a consequent loss of information.

$$\begin{array}{l}a \rightarrow_{R_1} \pi\\ \pi : b \rightarrow_{R_2} c\end{array} \quad \mid \quad \begin{array}{ccc}&a&\\ {\scriptstyle R_1}\swarrow&&\searrow{\scriptstyle R_2}\\ b&&c\end{array} \quad \mid \quad \begin{array}{l}\pi \rightarrow_{R_2} c\\ \pi : a \rightarrow_{R_1} b\end{array} \tag{3.3}$$

Each of the SDRSs above yields the same DG after simplification, namely, the one in the middle. The natural interpretation of a DG g describes the set of fully scoped SDRS structures that are compatible with these minimal requirements, *i.e* that would yield g by simplification. To get this set, every edge $r(x,y)$ in g, r, must be assigned left scope among the *descendants* of x in g (and right scope among those of y); this is a consequence of

i) x and y being *heads* of the left and right arguments of r and

ii) the SDRSs that are compatible with g do not admit relations with a right argument in one constituent and a left one outside of it.

Definition 15. Assume that we map each node[21] x of g into a unique variable $v_x \in V_l$ and each edge e into a unique variable symbol $r_e \in V_i$. Define $\bar{x}$ and $\bar{r}$ in an analogous way as in definition 14.

For a given dependency graph g, we compute a formula $\varphi_g = \exists\bar{x}\exists\bar{r}\ \psi_g \cup \Gamma_g$ such that

- ψ_g is defined analogously as in definition 14, defining the set of relation instances and EDUs.
- Γ_g is the formula stating the minimal scopes for each relation instance: for all edge in $e = R(x, y)$ in g, Γ_g entails i) r_e has v_x in its left scope and v_y in its right scope and ii) let $Des(x)$ be the set of variable symbols for all the descendants of x in g, Γ_g entails that if r_e has left scope over some v_z then v_z is in $Des(x)$ (symmetrically for y and right scope).

The interpretation $[\![g]\!]$ of a DG is: $\{\mathcal{M} \mid \mathcal{M} \vDash \varphi_g, A_0\text{-}A_3, A_6, A_7\}$. The DG $b \overset{R_1}{\leftarrow} a \overset{R_2}{\rightarrow} c$ for instance, is interpreted as a set of three structures isomorphic to the ones in (3.3) above.

We now relate DGs to RST Trees interpreted with the RNP. The idea is that the two kinds of structure express analoguous forms of dominance between coherence relations. However, DGs and their interpretation are generally slightly less constrained than RST Trees: for instance, given a DG g, we can always find RST Trees whose RNP interpretations are included in that of the DG, but not always an RST Tree which admits all of the DG interpretations at once under RNP. We therefore, focus on a restricted class of DGs, with a restricted set of interpretations.

3.4.2 Restrictions on DGs: Dependency Trees and the S_CDP$^+$ interpretation

A first restriction is needed because projectivity has different consequences for DGs and RST Trees: Assume that a relation r_1 holds between e_1 and e_2 and a relation r_2 holds between e_2 and e_3 with $e_1 <_t e_2 <_t e_3$. Assume furthermore that r_1 subordinates e_1 to e_2 and r_2 subordinates e_3 to e_1. In any projective graph representation of such a structure, the arc representing r_2 must be drawn "above" the arc for r_1. Now for RST with RNP this entails that r_1 cannot scope over e_3, while for a DG this is still permitted since r_2 makes e_3 a descendant of e_1 in the DG sense. Notice also that in the RST Tree, the "head" of the complex argument to r_2 is e_2 not e_1, which gives us an intuitive glance at why the RST Tree is not a faithful representation of the DG. This situation is illustrated, for instance, by figs. 3.1(a) and 3.1(c): in the RST Tree, the `Attribution` cannot take scope over C_4 whereas it can in the DG. Hence we must define the subset of DGs exempt of these non RST compatible dependencies, but also restrict the interpretation of DG to impose RST-like projectivity constraints on dominance. This is down below, after we expose the second restriction.

The second restriction concerns coordinating *Chain*s, and *Nest*s, as previously defined. Consider for instance a coordinating chains of 3 edus or more involving a single coordinating relation: $x_1 \rightarrow_{R_1} x_2 \rightarrow_{R_2} \cdots \rightarrow_{R_{n-1}} x_n$. We have, intuitively, two choices to represent such a structure in RST. Either we use a type 3 schema, or we binarize it. Neither solution is plainly satisfactory on its own: a type 3 schema will forbid that R_k takes right scope over any unit in x_{m+1} for $m > k$, whereas the original DG does. Binarising the *Chain* into a binary tree $R_1(x_1, R_2(x_2, \ldots, R_{n-1}(x_{n-1}, x_n))$,on the other hand, imposes by multinuclearity that each of the R_k scopes on the right over the whole set of nuclei of $x_{k+1}, x_{k+2}, x_{k+2} \ldots$, which accounts for only some of the possible interpretations of the DG. Ideally, the more general correspondence should be achieved by associating such a DG with the two elements set of RST Trees containing both versions,

[21] Recall that unlike RST Trees, DGs have EDUs as nodes and relations as edges.

with the type 3 schema, and with the binarized tree, thus covering all possible interpretations. An different but analoguous problem is encountered with *Nest*s when the unique nucleus does not textually precede or succede every satellites. However type 3 and 4 schemas, as they have been used so far in theoretical work and corpora cover only a restricted set of such problematic *Chain*s (namely those involving a single relation, like `Sequence`) and *Nest*s (actually very few, made of special combination of only 2 relations). Hence our choice here is to focus on binary RST Trees, without type 3 or 4 schemas, and establish a narrower correspondence by restricting the intepretation of coordinating *Chain*s to match that of their tree binarization in RST (under RNP). To this aim we introduce the interpretation principle S_CDP below. We remark nevertheless, that introducing more generic form of schemas into RST would make it possible to achieve larger correspondence.

Definition 16 ($\mathrm{DG}_{\mathrm{RST}}$, $[\![\cdot]\!]^{\mathrm{S\_CDP}^+}$)**.**

- A DG g is *RST-expressible* iff it is a DT (every EDU has at most one parent EDU) and for every arc $x \rightarrow^r y$, there is no unit z with $x <_t z <_t y$ and $z \rightarrow^{r'} x$. Let $\mathrm{DG}_{\mathrm{RST}}$ denote the set of RST-expressible DGs.
- S_CDP is an adapted and strengthened version of a principle called *Continuing Discourse Pattern*, (CDP, Asher and Lascarides, 2003). S_CDPstates that whenever a coordinating relation R_c^i originates as a node which appear to be also in the right scope of another relation R_s, R_s weakly dominate R_c^i on the right.
- $[\![\cdot]\!]^{\mathrm{S\_CDP}^+}$ is defined as the restriction of DG's interpretation $[\![\cdot]\!]$ conforming to S_CDP and futhermore verifying that whenever a relaion $x \rightarrow^r y$ holds in the DG, the corresponding relation weakly dominate on the left every relation holding between x and z such that $x <_t z <_t y$.

We have seen already that the DG on fig. 3.1(c) is not RST expressible because it involves $\texttt{Elaboration}(C_3, C_1)$ and the relation $\texttt{Attribution}(C_2, C_1)$. We can still apply the $\mathrm{S\_CDP}^+$ interpretation to it, which imposes that the external `Elaboration` dominates the `Attribution` on the left, hence to force the external `Elaboration` to scope over at least both C_1 and C_3 (since C_3 weakly dominates C_1).

We are now equipped with everything needed to formalize the correspondence.

3.4.3 Relation between DGs and RST

Overall this section, we consider only **binary** RST Trees. Hence, we will simply use the phrase "RST Trees" to refer to **binary** RST Trees.

Fact 3. We can define a translation $\mathcal{G}$ from RST Trees to dependency graphs, and a translation $\mathcal{H}$ from RST-expressible DGs to sets of RST Trees such that the following statements are true:

- For each RST Tree t, $\mathcal{G}(t)$ is an RST-expressible DG and $[\![t]\!]^{\mathrm{RNP}} \subseteq [\![\mathcal{G}(t)]\!]^{\mathrm{S\_CDP}^+}$.
- For each RST-expressible DG d $[\![d]\!]^{\mathrm{S\_CDP}^+} = \bigcup_{t \in \mathcal{H}(d)} [\![t]\!]^{\mathrm{S\_CDP}^+}$
- For each RST-expressible DG d, for each $t \in \mathcal{H}(d)$, $\mathcal{G}(t) = d$.

We provide the definition of $\mathcal{G}$ and $\mathcal{H}$ below:

Definition 17 (RST Trees to DGs)**.** The translation $\mathcal{G}$ takes a RST Tree t as input and outputs a pair $\langle G, n \rangle$, where $G = \langle Nodes, Edges \rangle$ is the image DG, and n an attachment point used along the recursive definition of $\mathcal{G}$.

- If t is an EDU x then $\mathcal{G}(t) = \langle (\{x\}, \{\}), x \rangle$.

- If $t = R(t_{1N}, t_{2S})$ then let $\langle G_1, n_1\rangle = \mathcal{G}(t_1)$ and $\langle G_2, n_2\rangle = \mathcal{G}(t_2)$.

$$\mathcal{G}(t) = \langle (G_1 \cup G_2 \cup \{R_{subord}(n_1, n_2)\}; n_1\rangle$$

- If $t = R(t_{1S}, t_{2N})$ then $\mathcal{G}(t) = \mathcal{G}(R(t_{2N}, t_{1S}))$
- If $t = R(t_{1N}, t_{2N})$ (multinuclear), let $\langle G_1, n_1\rangle = \mathcal{G}(t_1)$ and $\langle G_2, n_2\rangle = \mathcal{G}(t_2)$.

$$\mathcal{G}(t) = \langle (G_1 \cup G_2 \cup \{R_{coord}(n_1, n_2)\}; n_1\rangle$$

We can see DTs (which include RST-expressible DGs) as recursive structures with a root node a, and set of outgoing edges $r_o, \dots r_n$ linking a to the roots of dependent structures which are themselves DTs. We will therefore define $\mathcal{H}$ recursively. As we have seen $\mathcal{H}$ outputs a set of target RST Trees. This non-determinism comes from the possible choices in the order of processing of the relations $r_o, \dots r_n$.

To express these possible choices, and flesh out $\mathcal{H}$, we will need the notion of an inside-out strategy. Let a be an EDU and assume that a is part of a sequence

$$x_1 <_t \cdots <_t x_k <_t a <_t y_{k+1} <_t \cdots <_t y_{k+l}$$

of $k + l + 1$ textually ordonned EDUs. For such a sequence, an inside-out strategy centered at a, is a traversal of the $x's$ and y's in an order that respects the constraint that the x's are visited in increasing order of their distance to r, and that so are the y's. For instance, assuming $k = l$, both $x_k, x_{k-1}, \dots, x_2, x_1, y_{k+1}, y_{k+2}, \dots, y_{2k}$ and $x_k, y_{k+1}, x_{k-1}, y_{k+2}, \dots, x_{k-i}, y_{k+i+1}, \dots, x_1, y_{2k}$ are inside-out strategies. Formally, an inside-out strategy is a permutation[22] σ of the interval $[1; k + l]$ such that

$$\forall i \in [1, k + l]\ ((\sigma(i) \leq k \to \sigma^{-1}([\sigma(i); k]) \subseteq [1; i]) \wedge (\sigma(i) > k \to \sigma^{-1}([k + 1, \sigma(i)]) \subseteq [1; i])).$$

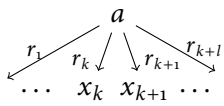
We are now able to define $\mathcal{H}$:

Definition 18 (From RST-expressible DGs to RST Trees)**.** Let d be an RST-expressible DG rooted in a (a is the unique consitutent is not the right argument of any relation). $\mathcal{H}$ is defined recursively:

- If d is an EDU, return $\mathcal{H}(d) = \{a\}$.
- If there is a single outgoing relation $a \to^R b$, let $d \restriction_b$ denote the subtree rooted in b and $T_b = \mathcal{H}(d')$.

$$\mathcal{H}(a) = \{R(a_{a_1}, t_{a_2}) \mid t \in T_b\}$$

  where a_1 and a_2 reflects the appropriate nuclearity type for R.

- If there are multiple relations 

then, assuming without loss of generality that $x_1 <_t \cdots <_t x_k <_t a <_t x_{k+1} <_t \cdots <_t x_{k+l}$, let σ be an inside-out strategy centered at a. Let $t_1, \dots, t_{k+1}$ be $k + 1$ RST Trees and let finally $\mathcal{H}_\sigma^{t_1,\dots t_n}(d)$ be defined as the output of the following procedure:

  1. Initialize an RST-Tree variable t with $t := a$.
  2. For i from 1 to k:

[22] in the mathematical sense of bijection of a set of integer onto itself

3 Select the relation $a \rightarrow^{r_{\sigma(i)}} x_{\sigma(i)}$. Applying $\mathcal{H}$ to the simple structure $a \rightarrow^{r_{\sigma}(i)} x_{\sigma(i)}$ results in a singleton set containing a unique (depth 1) RST Tree t'.

4 Replace a with t and $x_{\sigma(i)}$ with $t_{\sigma(i)}$ in t'. Update t with $t := t'$.

5 Output $\{t\}$.

Finally, let

$$\mathcal{H}(d) = \bigcup_{\substack{t_1 \dots t_{k+l} \in \mathcal{H}(d \restriction_{x_1}) \times \dots \times \mathcal{H}(d \restriction_{x_{k+l}}) \\ \sigma \text{inside-out strategy}}} \mathcal{H}_{\sigma}^{t_1,\dots,t_n}(d).$$

Recall the RST Tree (s_1) for example (3.1.1), also displayed on fig. 3.1(a). Applying $\mathcal{G}$ to this tree yields the following dependency tree (in linearized notation):

$$\texttt{Elaboration}_1(C_1, C_2) \wedge \texttt{Attribution}(C_3, C_1) \wedge \texttt{Elaboration}_2(C_3, C_4). \qquad (s_3')$$

Notice that it differs from the DG on fig. 3.1(c) which is **not** RST-Expressible. $[\![(s_3')]\!]^{\text{S\_CDP}^+}$ supports any RNP admissible reading of the RST Tree, but also an additional one:

$$\begin{array}{l} \texttt{Elaboration}_1([C_1], [C_2]) \\ \texttt{Elaboration}_2([C_3], [C_4]) \\ \texttt{Attribution}([C_1, C_2], [C_3, C_4]) \end{array} .$$

$\mathcal{H}((s_3'))$ yield in return two possible RST Trees (one for each of the two possible inside out strategies C_1, C_4 and C_4, C_1 at the top level), namely the orignal one (s_2), and a second one:

$$\texttt{Attribution}(\texttt{Elaboration}(C_{1N}, C_{2N})_S, \texttt{Elaboration}(C_{3N}, C_{4S})_N). \qquad (s_1')$$

This second tree indeed complete the possible interpretations with the one that was missing from (s_1')

$$[\![(s_3')]\!]^{\text{S\_CDP}^+} = [\![(s_1)]\!]^{RNP} \cup [\![(s_1')]\!]^{RNP}.$$

We have seen how we can use the scoping language and structures to formalise a common understanding of binary RST Trees and a fragment of Dependency Graphs. More generally, we can use the scoping language to compute a similarity score between any two scoping structures, which provides us a general mean of comparision for structures of the different theories and their different interpretations. We expose this in the next section.

3.5 Similarities and distances

The framework we have presented yields a notion of similarity that applies to structures of different formalisms. To motivate our idea, recall example 2.2.2; the structure in (s_3) in which `Attribution` just scopes over C_1 differs from the intuitively correct interpretation only in that `Attribution` should also scope over C_2 as in (s_2), while a structure that does this but in which C_3 is in the scope of the Elaboration relation is intuitively further away from the correct interpretation.

Our similarity measure *Sim* over structures $\mathcal{M}_1$ and $\mathcal{M}_2$ assumes a common set of elementary constituents and a correspondence between relation types in the structures. We measure similarity in terms of the scopes given to the relations. The intuition, is that given a map f from elements of relation instances in $\mathcal{M}_1$ to relation instances in $\mathcal{M}_2$, we achieve a similarity score by counting for each relation instance r the number of EDUs that are both in the left scope of one element of r and in $f(r)$, then divide this number

by the total number of diffrents constituents in the left scope of r_1 and r_2, and do the same for right scopes as well. The global similarity is given by the correspondence which yields the best score.

Given a relation $r_1 \in \mathcal{M}_1$ and a relation $r_2 \in \mathcal{M}_2$, let $\delta(r_1, r_2) = \begin{cases} 1 \text{ if } r_1 \text{ and } r_2 \text{ have the same label} \\ 0 \text{ otherwise} \end{cases}$.
Define $C_{\cdot|}(r_1, r_2) = |\{x : l \mid \mathcal{M}_1 \vDash x \in_{\cdot|} r_1 \wedge \mathcal{M}_2 \vDash x \in_{\cdot|} r_2\}|$, the number of constituents over which r_1 and r_2 scope and $D_{\cdot|}(r_1, r_2) = |\{x : l \mid \mathcal{M}_1 \vDash x \in_{\cdot|} r_1 \vee \mathcal{M}_2 \vDash x \in_{\cdot|} r_2\}|$. Define $C_{|\cdot}$ and $D_{|\cdot}$ analogously and assume that $\mathcal{M}_1$ has less relation instances than $\mathcal{M}_2$. Let $Inj(D_i^1, D_i^2)$ be the set of injections of relations instances of $\mathcal{M}_1$ to those of $\mathcal{M}_2$. Let $|\mathcal{M}|_i$ denotes the number of relations instances of a finite scoping structure $\mathcal{M}$.

$$Sim(\mathcal{M}_1, \mathcal{M}_2) = \frac{1}{2Max(|\mathcal{M}_1|_i, |\mathcal{M}_2|_i)} \times$$
$$\underset{f \in Inj(D_i^1, D_i^2)}{Max} \sum_{ri} \delta(r, f(r)) \times \left(\frac{C_{\cdot|}(r, f(r))}{D_{\cdot|}(r, f(r))} + \frac{C_{|\cdot}(r, f(r))}{D_{|\cdot}(r, f(r))}\right)$$

If M_2 has more relation instances, Invert arguments and use the definition above. If they have same number of instances, both directions coincide.

$$d(\mathcal{M}_1, \mathcal{M}_2) = 1 - Sim(\mathcal{M}_1, \mathcal{M}_2)$$

For a discourse structure $\mathcal{M}$, $Sim(\mathcal{M}, \mathcal{M}) = 1$; Sim ranges between 0 and 1. d is a Jaccard-like metric obeying symmetry, $d(x, y) = 0$ iff $y = x$, and d satisfies to the triangle inequality. One can further define the minimal, maximal or average similarity between any pair of structures of two sets S_1 and S_2. This gives an idea of the similarity between two underspecified interpretations, such as the ones provided by RNP of section 3.3.

For example, the maximal similarity between (s_2) interpreted as itself (immediate interpretation) and a scoping structure for the (general, non S_CDP$^+$) interpretation of the DG (s_3), is 7/12. It is achieved for the interpretation of (s_3) where `Attribution` is given left scope over C_1, C_2, C_4, $\texttt{Elaboration}_1$ holds between C_1 and C_2, and `Elaboration` fails to match the `Continuation` of (s_3). $max_{s \in [\![\zeta((s_2))]\!]} sim([\![(s_2)]\!], s) = 7/12$ also, because ζ must distribute $[2, 4]$ in (s_2) to avoid crossing dependencies; so $[\![\zeta((s_2))]\!] \cong [\![(s_3)]\!]$. The maximal similarity between (s_2) an interpretation of the RST tree in (s_1) with RNP (the same hold for (s_3')) is 19/36, achieved when both C_1 and C_2 are left argument of the `Attribution` (though not C_4, which isn't in any such interpretation). Under the MNP, the similarity drops to 17/36.

Given our results in section 3.4.1, we have:

Fact 4. i) For any RST-expressible DG g, for any S_CDP$^+$ valid interpretation of g, there an RST tree t and a RNP valid interpretation of t such that $Sim(g, t) = 1$.

ii) For any interpretation of an RST tree under RNP there is a DT g and an interpretation of g under S_CDP$^+$such that $Sim(t, g) = 1$.

We will now review some related work before concluding this chapter.

3.6 Related Work

Our work shares a motivation with Blackburn et al. (1993): Blackburn et al. (1993) provides a modal logic framework for formalizing syntactic structures; we have used MSO and our scope language to formalize discourse structures. While many concepts of discourse structure admit of a modal formalization, the fact

that discourse relations can have scope over multiple elementary nodes either in their first or second argument makes an MSO treatment more natural. Danlos (2008) compares RST, SDRT and Directed Acyclic Graphs (DAGs) in terms of their *strong generative capacity* in a study of structures and examples involving 3 EDUS. We do not consider generative capacity, but we have given a generic and general axiomatization of RST, SDRT and DT in a formal interpreted language. We can translate any structure of these theories into this language, independent of their linguistic realization. We agree with Danlos that the NP does not yield an accurate semantic representation of some discourses. We agree with Egg and Redeker (2010) that the NP is rather a constraint on structures, and we formalize this with the relaxed principle and show how it furnishes a translation from RS trees to sets of scoped structures. Danlos's interesting correspondence between restricted sets of RST trees, SDRSs and DAGs assumes an already fixed scope-interpretation for each kind of structure: SDRSs and DAGs are naturally interpreted as themselves, and RS Trees are interpreted with the mixed NP Our formalism allows us **both** to describe the structures themselves and various ways of computing alternate scopes for relations.

With regard to the discussion in Egg and Redeker (2008); Wolf and Gibson (2005b) of tree vs. graph structures, we show exactly how tree based structures like RST with or without the NP compare to graph based formalisms like SDRT. We have not investigated Graphbank here, but the scope language can axiomatize Graphbank (with A_0-A_3, A_8).

3.7 Conclusions and future directions

In this chapter, we have investigated the nature and expressive power of discourse structures. Our comparison is based on the idea that different kinds of structures denote, at the semantic level, different sets of admissible scopes for coherence relations. We have therefore explored in various formalisms the different mechanisms for computing this scopal information from the structures. We provided a unified account of these interpretative mechanisms, their input and target structures across theories using a pivot language and associated models to encode the scopes of relations regardless of theory-specific assumptions. We now dispose of a *lingua franca* for comparing discourse formalisms, helping us with a deeper thoeretical understanding of the different sorts of informations that different structures encode, and how these encodings relate. Importantly, we have shed a light on how some classes of structures (*e.g.* DGs, RST Trees) are interpreted as sets of (scoping structures isomorphic to) structures of other classes (*e.g.* SDRSs). Hence, we can see the former as underspecified representations of the latter. A measure of similarity between structures allows to quantify the semantic "deviation" between two different representations of the same discourse. Applied to a given "underspecified" structure, and one of its possible resolution, this similarity also quantities the information lost in the underspecification process. Our work may thus also be seen as contributing to clarify the relation between more syntax-oriented (introducing underspecification in their semantic interpretation) and more semantics-oriented (with transparent semantic interpration but less syntactically constrained, yielding larger search spaces in parsing tasks) views of discourse structure.

These theoretical results let us hope, in future work, for applications in the field of discourse parsing. One of our initial motivation was the sparcity of discourse annotated data, and we have made clear how to link different kind of annotated data, for instance how to switch from RST tree-like reprentations to Dependency Graphs; both formalisms have been used to train and evaluate discourse parsers, and an approach combining both sets of data seems a promising way to improve on existing models. Moreover, so-called *loss functions* play a significant role in models of structured learning (as *e.g.*, Crammer and Singer, 2003; Crammer et al., 2006) but a loss functions well-fitted for discourse is something hard to come up with. Our similarity measure takes a step in that direction. Also, the development of (sentencial) underspecification has played an important role in achieving wide-coverage grammar at the sentencial level, and we hope that the hierarchy of expressivity and the underspecified interpretations that we brought

up for the different discourse formalisms might furnish similar concepts and help the development of the syntax/semantics interface at the suprasentential level. Finally, our MSO formulation of structural constraints essentially involve constraints on graphs. Some of these constraints are expressible as constraints on undirected graphs, others use edges' directions, but should at least admit non-trivial relaxed undirected counterparts. This appeal to explore in future work automata-based techniques of model checking, such as the ones known form Courcelle's work (Courcelle and Engelfriet, 2012). Efficient (and particularly automata-based) model checking for our axioms is indeed something highly desirable for parsing methods based on constraint decoding such has Muller et al. (2012), where given a set of 'local' scores on structural fragments, a search is performed on the set of complete structures satisfying the target formalism's constraints to find a structure which globally maximises the sum of all of its fragments' scores.

The notion of "semantic deviation" that our similarity measure intends to capture is yet another field for further investigations and the target of the next chapter: the similarity we proposed, allthough its comparison of the full semantic scopes of discourse-level logical operators makes it less sensible to small differences in syntactic boundaries than other, more classical, distances defined over syntactic structures such as ParseEval (Black et al., 1991) and Leaf Ancestor (Sampson, 2000), might still be objectable as a candidate for measuring "semantic" similarity: one can object, first, that it compares only different structures built over the same set of EDUs and, second, that in the end, the comparison is still essentially one of syntactic boundaries, even though it relies on more fine-grained syntactic primitives. For instance, matching an `Elaboration` with an `Explanation` between two structures will weight in the exact same way as matching an `Elaboration` with a `Contrast`: it will count as as failure and yield a score of 0. Yet an `Elaboration` is intuitively semantically closer to an `Explanation` than to a `Contrast`.

This illustrates the need to investigate what one could call "essentially semantic" metrics. What exactly constitutes an essentially semantic metric is an interesting problem in its own right. Intuitively, it should strongly rely on the model theoretic interpretation of structures, so that, in particular, structures expressing the same "meaning", even realized differently syntactically, are judged very close to each other. Efforts have been made in that direction to characterise logically equivalent structures (*e.g.*, Roze, 2013; Asher et al., 2011), but do not propose definitions of metrics or similarity functions. On another note, such metrics should equally importantly accounts for the "dynamics" of discourse and dialogue, *i.e.* predicting how and when these grow closer or further away as new information comes into play.

The next chapter is concerned with these issues and proposes a fundational studies of *semantic metrics*: what they should be, which axioms they should satisfy, and which functions are good candidate for semantic metrics. Toward this aim, we switch from the notion of discourse discussed so far to a broader and more abstract notion of *conversation*. Conversations in that sense, are simply sequences of token of a given vocabulary. Working at this high level of abstraction will allow us to investigate in the simplest setting and without presuming too much of the nature of conversations, the respective structural properties that conversations, their interpretation, and semantic distances should have with respect to one another. Parts II and III on the other hand will review and propose accounts of more concrete implementation of the notions of "dialogue" and "conversation".

Chapter 4

Semantic distances

Contents

4.1 Introduction

If linguistic behavior is to be analyzed as a form of rational behavior (Grice, 1967), it is important to be able to assess the conversational goals of linguistic agents and the extent to which they are fulfilled by any given conversation in a manageable way. Specifying preferences over the set of all possible choices of what to say is clearly intractable for us as theorists and for speakers as practioners. Instead, speakers must be able to group conversations into semantically similar classes and to assess the relative semantic proximity of any two pairs of conversations. The preferences of the agents over different ways of expressing themselves have to do with how close these ways are from satisfying certain positive or negative semantic goals. An elegant way to be able to do this, is to have a metric over conversations that is semantic in nature. The goal of this chapter is to identify properties that characterize 'semantic metrics' and to identify reasonable axioms that can help us isolate well-behaved semantic metrics.

A workable definition of semantic distance between texts or conversations is also important for the evaluation of annotations of discourse structure in text and dialogue. It is also crucial to the success of the machine learning of semantic structures from annotated data, as all known algorithms rely on some notion of similarity or loss with respect to the target structure. While measures of syntactic similarity like ParseEval (Black et al., 1991) and Leaf Ancestor (Sampson, 2000) are well-understood and used in computational linguistics, they yield intuitively wrong results. ParseEval, for instance, places too much importance on the boundaries of discourse constituents, which are often notoriously hard even for expert annotators to agree on. Investigations of distances between semantic interpretations of a text are rarely examined. While a natural equivalence and ordering relation over contents comes from the underlying logic of formal semantic analysis, this only gives a very crude measure. Some have appealed to a language of semantic primitives to exploit the more developed measures of syntactic distance in a more semantic setting. But such an approach depends on the choice of semantic primitives, with no clear consensus on how to go about determining these primitives.

Semantic distances are also relevant in the context of formal theories of belief revision. Lehmann et al. (2001) explores Alchourrón et al. (1985) style postulates that characterize a wide family of belief revision operator based on pseudo-distances on models satisfying only very mild assumptions. Our problem is also closely related to the problem of determining the distance of a scientific theory from the truth. This problem, referred to as the problem of verisimilitude or truthlikeness in philosophy of science (since Popper (1968)), is arguably reducible to the problem of having a satisfactory concept of similarity between theories in a formal language.

The aim of this chapter is to study semantic metrics for an abstract and simple concept of conversations. Syntactically, we assume that conversations are monoids with respect to concatenation. These conversations are equipped with an interpretation function mapping them into some distinct semantic space. In general, our assumptions about the semantic space and the interpretation function will be as minimal as possible. As far as identifying the axioms that characterize our concept of 'semanticity' for a metric goes, we will not be making any assumption. To analyze candidate axioms that characterize *well-behaved* semantic metrics, it will be interesting to consider the effect of assuming a bit more structure. Specifically we will pay some attention to the case in which the semantic co-domain of the interpretation function is a lattice. As an example, sequences of propositional formulas with their classical interpretation certainly fall under this category. We will moreover consider interpretation functions that satisfy some structural properties, for example assuming that the semantic meaning of a sequence is invariant under stuttering, that is immediate repetition of the same element in a sequence, or even assuming complete invariance under permutation.

To develop semantic metrics for conversations in natural language or for their representations in some formalism suitable for discourse interpretation (like for instance SDRT) we first need to clarify the space of reasonable axioms and metrics for the simplest and most general representations. We take this first step here.

The chapter is organized as follows. We start with a first section, section 4.2, that contains technical preliminaries and settles the notation. We then describe in section 4.3 different properties that can be met by an interpretation function, regarding how it interacts with sequences-concatenation or the structure of the semantic space. Section 4.4 introduces some elementary background about generalized metrics and metrics over subsets of a metric space. Section 4.5 draws a map of different level of *semanticity* for metrics, corresponding to different requirements on the interaction between the interpretation function and the metric. Section 4.6 introduces some concrete candidate semantic metrics. Sections 4.6 to 4.8 then describe how these potential measures of semantic similarity fare with respect to different lists of axioms. We show that certain combinations of axioms lead to trivialization results. We then conclude.

4.2 Preliminaries and notation

This section contains some technical preliminaries and settles notation. The reader can skip this section on a first reading, and come back to it when needed.

4.2.1 Sets, functions, sequences, orders and lattices

Let X, Y be two sets. We let $X \ominus Y$ denote the symmetric difference of X and Y. Let $card(X)$ denote the cardinality of X. let X^* be the set of finite strings over X. If $f : X \to Y$, we let $f(X)$ and $f[X]$ be alternative notation for the image of X under f, that is $f(X) = f[X] = \{f(x) | x \in X\}$. We let $dom(f) = X$, $target(f) = Y$ and $ran(f) = f(X)$. The *kernel* of f is the equivalence relation $\sim$, such that $x \sim y$ iff $f(x) = f(y)$. We say that f is *isotone* whenever for all x, y with $x \leq y$ we have $f(x) \leq f(y)$. We write $f : X \hookrightarrow Y$ whenever f is partial function from X to Y, that is there exists a non-empty subset of $A \subseteq Y$ such that $f : A \to Y$.

Given a sequence $\sigma \in X^*$ we let $len(\sigma)$ be the length of σ. If $k \leq len(\sigma)$ then we let $\sigma|_k$, be the prefix of σ of length k, and we let $\sigma(k)$ or $\sigma[k]$ be the k^{th} element of σ. We let $ran(\sigma)$ be the range of σ, that is $ran(\sigma) = \{\sigma[k] | 1 \leq k \leq len(\sigma)\}$. We let $\overrightarrow{\epsilon}$ be the empty sequence.

A relation $\leq$ on X is a *pre-order* on X iff it is a reflexive and transitive relation on X. $(X, \leq)$ is a *poset* iff $\leq$ is a pre-order on X such that $\leq$ is antisymmetric on X, that is $x \leq y$ and $y \leq x$, implies that $x = y$. A *lattice* $(X, \leq)$ is a poset such that every two elements $x, y \in X$ have a least upper bound (or join, denoted $x \wedge y$) and and a greatest lower bound (or meet, denoted $x \vee y$). A lattice $(X, \leq)$ is bounded whenever X has a least and a greatest element (denoted $\bot$ and $\top$).

4.2.2 Propositional Languages and interpretation functions

Given a language L we let $\varphi, \psi, \chi, \varphi_1, \varphi_2 \ldots$ range over L, and we let $\overrightarrow{\varphi}, \overrightarrow{\psi}, \overrightarrow{\sigma}, \overrightarrow{\tau}, \overrightarrow{\upsilon}, \overrightarrow{\sigma_1}, \overrightarrow{\sigma_2} \ldots$ range over L^*. Given a finite set $\textsc{prop} = \{p_1, \ldots, p_n\}$ we let $\mathsf{L}_{\textsc{prop}}(1)$ be defined as follows:

$$\varphi ::= p | \neg p | \top | \bot$$

where p ranges over $\textsc{prop}$. And we define $\mathsf{L}_{\textsc{prop}}$ as follows:

$$\varphi ::= p | \top | \bot | \neg \varphi | \varphi \wedge \varphi | \varphi \vee \varphi$$

where p ranges over $\textsc{prop}$. We define $sig(\varphi) : \mathsf{L} \to \wp(\textsc{prop})$ where $p \in sig(\varphi)$ iff p occurs in φ. Given a sequence $\overrightarrow{\sigma} \in \mathsf{L}^*$ or a subset $A \subseteq \mathsf{L}$, we write $sig(\overrightarrow{\sigma}) := \bigcup_{\psi \in ran(\overrightarrow{\sigma})} sig(\psi)$ or $sig(A) := \bigcup_{\psi \in A} sig(\psi)$, respectively.

Classical truth-functional interpretation of $\mathsf{L}_{\textsc{prop}}$. Let $W_{\textsc{prop}} = 2^{\textsc{prop}}$. Depending on context, will treat a member $V \in W_{\textsc{prop}}$ either as a function $V : \textsc{prop} \to \{0, 1\}$ or as a subset $V \subseteq \textsc{prop}$. These

two representations are of course equivalent. We let $[\![\cdot]\!]^t : \mathsf{L} \to 2^{\mathrm{PROP}}$, be the classical truth-functional interpretation function of $\mathsf{L}_{\mathrm{PROP}}(1)$ and $\mathsf{L}_{\mathrm{PROP}}$.

4.3 Properties of interpretation functions

In general, we will work with abstract concepts of a language and of an interpretation function. Let L, X be non-empty sets.

Definition 19 (Interpretation function)**.** An *interpretation function* of L into X is a function $\|\cdot\| : \mathsf{L}^* \to X$.

An *interpretation function* for L is an interpretation function L into Y for some non-empty set Y.

4.3.1 Co-domain

In this paper, we will sometimes assume that the semantic space has some structure. We are always explicit about these assumptions whenever we make them.

Definition 20. We say that $\|\cdot\|$ is $(W,\leq)$*–pre-order-valued* ($(W,\leq)$*–poset-valued*) whenever *target*$(\|\cdot\|)=W$ and $(W,\leq)$ is a pre-order (respectively, a poset).

We say that $\|\cdot\|$ is *pre-order–valued*, iff it is $(W,\leq)$*–pre-order-valued* for some pre-ordered set $(W,\leq)$, and similarly for *poset–valued*.

Definition 21. $\|\cdot\|$ is $(W,\leq,\wedge,\vee)$*–lattice-valued* iff *target*$(\|\cdot\|) = W$ and $(W,\leq)$ is a lattice, with $\wedge$ and $\vee$ as its *meet* and *join* operator, respectively.

Definition 22. $\|\cdot\|$ is $(W,\leq,\wedge,\vee,\bot,\top)$*–lattice-valued* iff $\|\cdot\|$ is $(W,\leq,\wedge,\vee)$*–lattice-valued*, and $(W,\leq)$ is a bounded lattice, with $\bot$ and $\top$ as its least and greatest element, respectively.

We say that $\|\cdot\|$ is *lattice-valued*, iff it is $(W,\leq,\wedge,\vee)$*–lattice-valued* for some $(W,\leq)$, $\wedge$ and $\vee$. We say that $\|\cdot\|$ is *bounded lattice-valued*, iff it is $(W,\leq,\wedge,\vee,\bot,\top)$*–lattice-valued*, for some $(W,\leq)$, $\wedge$, $\vee$, $\bot$ and $\top$.

Definition 23. $\|\cdot\|$ is *set-valued* iff we have *target*$(\|\cdot\|) = \wp(W)$ for some non-empty set W.

4.3.2 Structural properties for interpretation functions

It will sometimes be interesting to restrict ourselves to interpretation functions satisfying certain structure properties. Assume that $\|\cdot\|$ is $\leq$*–poset-valued*. Below the comma ',' is the concatenation operator. The axiom in the table below, are to be understood as quantifying universally. $\vec{\alpha}$, $\vec{\beta}$ ranging over $dom(\|\cdot\|)^*$ and α, β ranging over $dom(\|\cdot\|)$.

Axiom name	Meaning
contraction	$\|\vec{\alpha}, \varphi, \varphi, \vec{\beta}\| \leq \|\vec{\alpha}, \varphi, \vec{\beta}\|$
expansion	$\|\vec{\alpha}, \varphi, \vec{\beta}\| \leq \|\vec{\alpha}, \varphi, \varphi, \vec{\beta}\|$
exchange	$\|\vec{\alpha}, \varphi, \psi, \vec{\beta}\| = \|\vec{\alpha}, \psi, \varphi, \vec{\beta}\|$
right monotonicity	$\|\vec{\alpha}, \varphi\| \leq \|\vec{\alpha}\|$
left monotonicity	$\|\varphi, \vec{\alpha}\| \leq \|\vec{\alpha}\|$
$\vec{\epsilon}$–$\top$	$\|\vec{\alpha}\| \leq \|\vec{\epsilon}\|$
adjunction	If $\|\vec{\alpha}\| \leq \|\vec{\beta}\|$ and $\|\vec{\alpha}\| \leq \|\vec{\gamma}\|$ then $\|\vec{\alpha}\| \leq \|\vec{\beta}\,\vec{\gamma}\|$
mix	If $\|\vec{\alpha}_1\| \leq \|\vec{\beta}_1\|$ and $\|\vec{\alpha}_2\| \leq \|\vec{\beta}_2\|$ then $\|\vec{\alpha}_1\,\vec{\alpha}_2\| \leq \|\vec{\beta}_1\,\vec{\beta}_2\|$

For example, $\|\cdot\|$ satisfies contraction iff for every $\vec{\alpha}, \vec{\beta} \in \mathsf{L}^*$ and $\varphi \in \mathsf{L}$ we have $\|\vec{\alpha}, \varphi, \varphi, \vec{\beta}\| \leq \|\vec{\alpha}, \varphi, \vec{\beta}\|$.

Remark 1. If $\|\cdot\|$ satisfies exchange, then $\|\cdot\|$ for every $\vec{\varphi}$ and $\vec{\psi}$ that are equivalent up to permutation we have $\|\vec{\varphi}\| = \|\vec{\psi}\|$. If $\|\cdot\|$ satisfies either right or left monotonicity, then $\|\cdot\|$ satisfies $\vec{\epsilon} - \top$.

4.3.3 Stronger properties

Definition 24 (Conjunctive, intersective intepretation)**.**

- $\|\cdot\|$ is conjunctive iff it is lattice-valued and $\forall \vec{\varphi}, \vec{\psi}, \|\vec{\varphi}\,\vec{\psi}\| = \|\vec{\varphi}\| \wedge \|\vec{\psi}\|$.
- $\|\cdot\|$ is intersective iff it is set-valued and $\forall \vec{\varphi}, \vec{\psi}, \|\vec{\varphi}\,\vec{\psi}\| = \|\vec{\varphi}\| \cap \|\vec{\psi}\|$.

Definition 25. We let $\|\cdot\|^t$ be the interpretation function for L^* defined by $\|\vec{\varphi}\|^t := [\![\bigwedge_{\varphi \in ran(\vec{\varphi})}]\!]^t$.

Example 1. $\|\cdot\|^t$ is intersective. If $\|\cdot\|$ is intersective, then it is bounded $\subseteq$–lattice-valued.

4.4 Generalized metrics

We start by recalling some basic definitions.

Definition 26 (Semi-Pseudometric)**.** A *semi-pseudometric* on a set X is a function $d : (X \times X) \to \mathbb{R}$, such that for all $x, y, z \in X$ we have:

1. $d(x, x) = 0$;
2. $d(x, y) = d(y, x)$.

Definition 27 (Pseudometric)**.** A *pseudometric* on a set X is a semi-pseudometric on X, such that for all $x, y, z \in X$ we have:

3. $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality).

If d is a (semi-)pseudometric on X, then (X, d) is a (semi-)pseudometric space.

Definition 28 (Trivial pseudo-metric)**.** The *trivial pseudo-metric* over a set A is the function

$$d : \begin{cases} A \times A \to \mathbb{R} \\ \forall x, y \in A \quad d(x, y) = 0 \end{cases}$$

4.4.1 Metrics on valuations, relations and graphs

Given a finite set PROP Hamming distance on 2^{PROP} is the metric $\delta_{ham} : 2^{\text{PROP}} \times 2^{\text{PROP}} \to \omega$ defined as $\delta_{ham}(V, V') = card(V \ominus V')$.

4.4.2 Aggregators

Let δ_i be a pseudo-metric on a set X. We want to study closeness between subsets of X, and so we provide some natural aggregators α associating with δ_i a function $d^i_\alpha : 2^X \times 2^X \to \mathbb{R}$, that may or may not be a pseudo-metric, depending on the particular aggregator.

Definition 29 (min aggregator)**.** Let $d^i_{\min}(A, B) = \min_{x \in A, y \in B} d_i(x, y)$.

Definition 30 (max aggregator)**.** Let $d^i_{\max}(A, B) = \max_{x \in A, y \in B} d_i(x, y)$.

Definition 31 (*Hausdorff* aggregator)**.** Formally $d^i_H(A,B) = \max\{\max_{x\in A}\min_{y\in B} d_i(x,y), \max_{y\in B}\min_{x\in A} d_i(x,y)\}$.

Definition 32 (*mean* aggregator)**.** Formally $d^i_{am}(A,B) = \sum_{x\in A, y\in B} \frac{1}{card(A\times B)} d_i(x,y)$.

Remark 2. In general max and *mean* will return a non-zero value for (A,A). Note also that the min aggregator will return 0 for (A,B) whenever $A\cap B \neq \varnothing$.

Let W be a set and let d be a pseudo-metric on W. Let L be a language and let $\|\cdot\|$ be an interpretation function for L such that $target(\|\cdot\|) = \wp(W)$ for some non-empty set W. Let $d^i_\alpha : \wp(W)\times\wp(W)\to\mathbb{R}$ be an aggregator based on the distance d_i between points of W. We let $d^i_{\alpha,\mathsf{L},\|\cdot\|} : \mathsf{L}^*\times\mathsf{L}^*\to\mathbb{R}$ be defined by $d^i_{\alpha,\mathsf{L},\|\cdot\|}(\overrightarrow{\varphi},\overrightarrow{\psi}) = d^i_\alpha(\|\overrightarrow{\varphi}\|,\|\overrightarrow{\psi}\|)$. When L and $\|\cdot\|$ are clear from context, we will simply write d^i_α for $d^i_{\alpha,\mathsf{L},\|\cdot\|}$. For instance, $d^{ham}_{H,\mathsf{L}^{prop},\|\cdot\|^t}$ is sometimes shortened as d^{ham}_H when L^{prop} and $\|\cdot\|^t$ are clear from context.

4.5 What is a semantic metric?

Now that we have set the stage for our investigations, our first task is to define our object of interest: semantic pseudometrics. Semantic pseudometrics are a subclass of linguistic pseudometrics.

Definition 33 (Linguistic (semi-)pseudometric)**.** A *linguistic* (semi-)pseudometric on a language L is a partial function $d : (\mathsf{L}^*\times\mathsf{L}^*)\hookrightarrow\mathbb{R}$ such that $dom(d)$ is symmetric and $(dom(d), d)$ is a (semi-)pseudometric space.

How semantic pseudometrics should be defined is not a fully straightforward matter. A minimal requirement would be the following:

$$\begin{aligned} &\text{If for every } \overrightarrow{\chi_1},\overrightarrow{\chi_2} \text{ with } \|\overrightarrow{\chi_1}\| = \|\overrightarrow{\chi_2}\| \\ &\text{we have } d(\overrightarrow{\varphi},\overrightarrow{\chi_1}) = d(\overrightarrow{\psi},\overrightarrow{\chi_2}) \quad \text{then} \quad \|\overrightarrow{\varphi}\| = \|\overrightarrow{\psi}\| \end{aligned} \tag{min sem separation}$$

That is, if two sequences $\overrightarrow{\chi_1},\overrightarrow{\chi_2}$ are semantically non-equivalent, then there should be two (other) semantically equivalent sequences, that are not pairwise equidistant from $\overrightarrow{\chi_1}$ and $\overrightarrow{\chi_2}$. A stronger, yet reasonable, assumption is:

$$\text{If for every } \overrightarrow{\chi} \text{ we have } d(\overrightarrow{\varphi},\overrightarrow{\chi}) = d(\overrightarrow{\psi},\overrightarrow{\chi}) \quad \text{then} \quad \|\overrightarrow{\varphi}\| = \|\overrightarrow{\psi}\| \tag{sem separation}$$

The axiom states that if two sequences of formulas $\overrightarrow{\varphi}$ and $\overrightarrow{\psi}$ are not semantically equivalent, then there is some sequence of formulas $\overrightarrow{\chi}$ that is not at the same distance from both $\overrightarrow{\varphi}$ and $\overrightarrow{\psi}$.

Fact 1. Let d be a semi-pseudometric. If d satisfies (sem separation), then it satisfies (min sem separation).

Finally we consider a stronger axiom:

$$\text{If } d(\overrightarrow{\varphi},\overrightarrow{\psi}) = 0 \quad \text{then} \quad \|\overrightarrow{\varphi}\| = \|\overrightarrow{\psi}\| \tag{zero $\Rightarrow$ sem$\equiv$}$$

The axiom is a regularity condition stating, that semantically non-equivalent sequences of formulas, should be at positive distance of each other.

Fact 2. Let d be a semi-pseudometric. If d satisfies (zero $\Rightarrow$ sem$\equiv$), then it satisfies (sem separation).

The two become equivalent if we assume triangle inequality.

Fact 3. Let d be a pseudo-metric. d satisfies (sem separation) iff d satisfies (zero $\Rightarrow$ sem$\equiv$).

The converse of (zero ⇒ sem≡), below, states that semantically equivalent sequences formulas, should be a distance 0 of each other.

$$\text{If } \|\vec{\varphi}\| = \|\vec{\psi}\| \quad \text{then} \quad d(\vec{\varphi}, \vec{\psi}) = 0 \qquad \text{(sem≡ ⇒ zero)}$$

Unsurprisingly (sem≡ ⇒ zero) will filter out syntactically driven notions such as δ_{count} or $\delta_{synt,count}$.

Definition 34. Given a language L, let $\delta_{count}(\vec{\varphi}, \vec{\psi}) := card(ran(\vec{\varphi}) \ominus ran(\vec{\psi}))$

Definition 35. Given a language L, let $\delta_{synt,count}(\vec{\varphi}, \vec{\psi}) := \delta_{count}(\vec{\varphi}, \vec{\psi}) + card(sig(\vec{\varphi}) \ominus sig(\vec{\psi}))$

Fact 4. $(\mathsf{L}_{\text{PROP}}(1), \delta_{count})$ does not satisfy (sem≡ ⇒ zero).

Fact 5. $(\mathsf{L}_{\text{PROP}}(1), \delta_{synt,count})$ does not satisfy (sem≡ ⇒ zero).

As observed previously, (sem≡ ⇒ zero) rules out a number of aggregators, e.g. :

Fact 6. $(\mathsf{L}_{\text{PROP}}, d^{ham}_{\max})$ does not satisfy (sem≡ ⇒ zero).

We have seen that (sem≡ ⇒ zero) and the triangle inequality together imply that two semantically equivalent points are equidistant to any other third point. This latter notion of semantic invariance implies in return (sem≡ ⇒ zero) and might be a desirable property as well:

$$\text{If } \|\vec{\varphi}\| = \|\vec{\psi}\| \quad \text{then} \quad \text{for every } \vec{\chi} \text{ we have } d(\vec{\varphi}, \vec{\chi}) = d(\vec{\psi}, \vec{\chi}) \qquad \text{(sem preservation)}$$

Fact 7. Let d be a semi-pseudometric. If d satisfies (sem preservation) then d satisfies (sem≡ ⇒ zero).

Finally, we can require our (semi-)pseudometric to be fully induced by a distance on the co-domain of the interpretation function $\|\cdot\|$, which we define as follows:

$$\text{If } \|\vec{\varphi_1}\| = \|\vec{\psi_1}\| \text{ and } \|\vec{\varphi_2}\| = \|\vec{\psi_2}\| \quad \text{then} \quad d(\vec{\varphi_1}, \vec{\varphi_2}) = d(\vec{\psi_1}, \vec{\psi_2}) \qquad \text{(sem induced)}$$

Fact 8. Let d be a semi-pseudometric. If d satisfies (sem induced) then d satisfies (sem preservation).

Fact 9. Let d be a pseudo-metric. If d satisfies (sem≡ ⇒ zero), then it satisfies (sem induced).

Corollary 1. Let d be a pseudo-metric that satisfies (sem≡ ⇒ zero). There exists a pseudo-metric $\dot{d}$ on $ran(\|\cdot\|)$ such that $\dot{d}(\|\vec{\varphi}\|, \|\vec{\psi}\|) = d(\vec{\varphi}, \vec{\psi})$.

Fact 10. Let d be a pseudo-metric that verifies (sem≡ ⇒ zero). Let ≡ be the kernel of $\|\cdot\|$. The following holds:

1. If $\vec{\varphi} \equiv \vec{\psi}$ then $\forall \vec{\chi}\ d(\vec{\chi}, \vec{\varphi}) = d(\vec{\chi}, \vec{\psi})$.

Fact 11. Let d be a pseudo-metric. If d satisfies (min sem separation) and (sem≡ ⇒ zero), then it satisfies (sem separation).

Figure 4.1, on p.62, summarizes the relation between the axioms discussed in this section. We are now ready to define our notion of 'semanticity'.

Definition 36 (Semantic Pseudometric)**.** A linguistic (semi-)pseudometric is semantic whenever for all $\vec{\varphi}, \vec{\psi}$ we have $\|\vec{\varphi}\| = \|\vec{\psi}\|$ iff for all $\vec{\chi}_1, \vec{\chi}_2$ such that $\|\vec{\chi}_1\| = \|\vec{\chi}_2\|$ we have $d(\vec{\varphi}, \vec{\chi}_1) = d(\vec{\psi}, \vec{\chi}_2)$.

Fact 12. A linguistic semi-pseudometric is semantic iff it satisfies (min sem separation) and (sem induced).

Fact 13. A linguistic pseudometric is semantic iff it satisfies (min sem separation) and (sem≡ ⇒ zero).

Now that we have settled our definition of semantic pseudo metric, which we will use in the sequel, we can tackle our main problem—in brief:
What are reasonable properties of a semantic pseudometric on (a subset of) L^*?

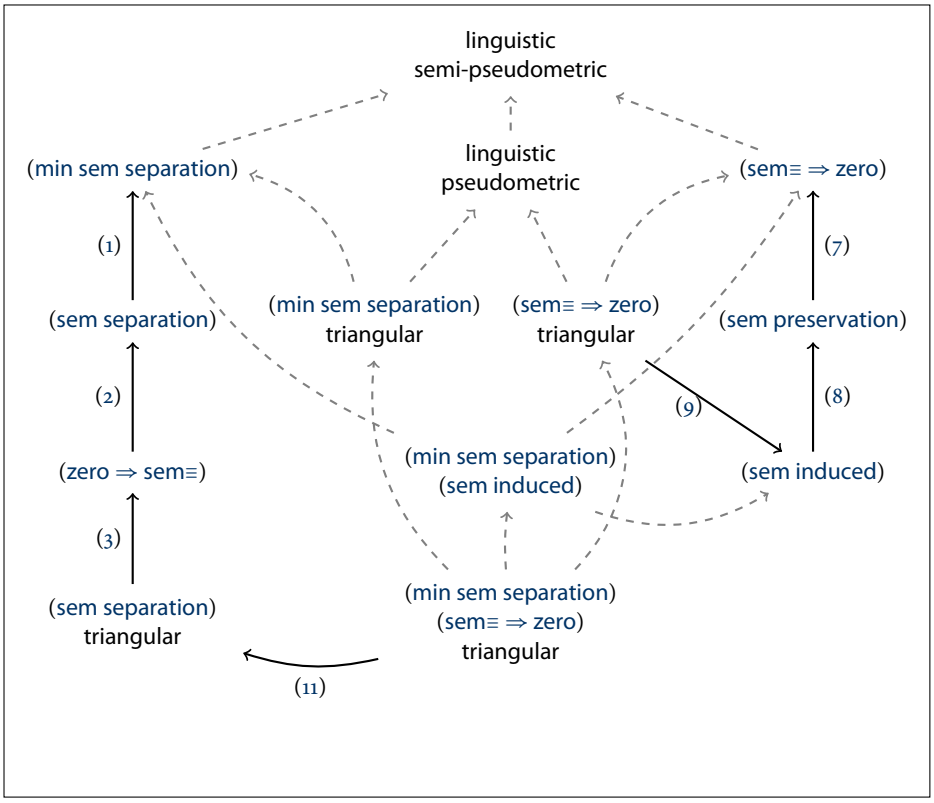


Figure 4.1: Summary of the results in Section 4.5. Dashed arrows follow from definitions.

4.6 Axioms for semantic metrics

In this section we will investigate how various semantic pseudo metrics behave with respect to some different axioms, some more intuitive than others. We have already introduced the Hausdorff metric d_H^{ham} over sequences of formulas in L_{PROP}. We now provide some other metrics.

4.6.1 Examples of semantic metrics

If $\|\cdot\|$ is finite set-valued, we can define the following metrics:

Definition 37 (Semantic Symmetric Difference Metric)**.** The semantic symmetric difference metric is the cardinality of the symmetric difference between the respective interpretations. $d_\ominus(\vec{\varphi}, \vec{\psi}) := card(\|\vec{\varphi}\| \ominus \|\vec{\psi}\|)$

Definition 38 (Proportional metric)**.** The proportional metric decreases from 1 to 0 as the ratio between the intersection and the union of the respective interpretations increases.

$$d_\alpha(\vec{\varphi}, \vec{\psi}) = \begin{cases} 0 & \text{if } \|\vec{\varphi}\| = \|\vec{\psi}\| = \varnothing \\ 1 - \frac{card(\|\vec{\varphi}\| \cap \|\vec{\psi}\|)}{card(\|\vec{\varphi}\| \cup \|\vec{\psi}\|)} & \text{otherwise} \end{cases}$$

Observe that, with this measure, pairs of sequences with non-empty disjoint interpretations will always be at distance 1 of each other.

The idea behind the next metric is to measure closeness by counting how many possible *continuations* one sequence has that can not apply to the other. The motivation is twofold: first, it follows the same line of thought that lies behind $d_\ominus$ or d_α (which is, roughly, to measure how big the set of semantic values that "separates" two sequences is), but seems more likely applicable to semantic spaces differing from lattices of sets –for instance, the kind of spaces used in dynamic semantics. As we outlined in previous chapters, there are several fundamental differences between "dynamic" interpretation functions and classical "static" ones. Dynamic cunjunction ($\wedge_D$) is not commutative and dynamic entailment is not always reflexive. In particular, $\varphi \wedge_D \psi$ does **not** coincide with the meet of the dynamic-entailment preorder (which is in fact captured by $\neg(\neg\varphi \vee_D \neg\psi)$). Hence, dynamic interpretation is not cunjunctive in our terms. Nevertheless, admissible continuations should naturally adapt to the dynamic case: in DPL for instance, understanding "admissible" as "consistent", $\neg P(x)$ can be an admissible continuation of $\neg\neg\exists x P(x)$ whereas it is never one of $\exists x P(x)$. In SDRT, further assuming the right frontier constraint as part of admissibility, $\texttt{Elaboration}(\pi_2, \pi_3)$ is an admissible continuation of $\texttt{Explanation}(\pi_1, \pi_2)$ while it is not one of $\texttt{Result}(\pi_2, \pi_1)$. Yet many other continuations will be admissible in both cases, so that a continuation-based metric should separate the two conversations but still judge them quite close to each other. Notions such as symmetric difference seem to the least less straightforward to adapt to a dynamic setting in such an intuitive way.

But then again, subsequent sections will provide a second motivation: it turns out that even in a simple set-valued setting, with a cunjunctive interpretation, the continuation-based metric fits some intuitions (namely (weak $\wedge$ rule)) better than the other metrics explored so far. This is exposed in section 4.8.

For a set-valued interpretation function, we provide a formalization of a continuation-based pseudometric. We give the definition below, then explain why it is indeed a continuation-based pseudometric:

$$d_C(\overrightarrow{\varphi}, \overrightarrow{\psi}) = \frac{2^{card(\|\overrightarrow{\varphi}\|)} + 2^{card(\|\overrightarrow{\psi}\|)} - 2 \cdot 2^{card(\|\overrightarrow{\varphi}\| \cap \|\overrightarrow{\psi}\|)}}{2^{card(\|\overrightarrow{\varphi}\|)+card(\|\overrightarrow{\psi}\|)}}.$$

This pseudometric is continuation-based because in the simplest setting of building a semantic metric for $\mathsf{L}^*_{\textsc{prop}}$ with the classical truth functional $\|\cdot\|^t$ we can derive the above expression by defining an admissible continuation as a *consistent* one: a set $\|\overrightarrow{\chi}\|^t$ is an admissible continuation of $\overrightarrow{\varphi}$ just in case $\|\overrightarrow{\varphi}\,\overrightarrow{\chi}\|^t = \|\overrightarrow{\varphi}\|^t \cap \|\overrightarrow{\chi}\|^t \neq \varnothing$. Since, in return, for every subset $S \subseteq \|\overrightarrow{\varphi}\|^t$ there exists a $\chi \in \mathsf{L}_{\textsc{prop}}$ such that $\|\chi\|^t = S$, an admissible continuation for $\overrightarrow{\varphi}$ is, up to semantic equivalence, a set of semantic values that has non-empty intersection with $\|\overrightarrow{\varphi}\|^t$. Then, an admissible contibuation for $\overrightarrow{\varphi}$ which is not admissible by $\overrightarrow{\psi}$ is, up to semantic equivalence, a set of semantic values S such that $S \cap \|\varphi\|^t \neq \varnothing$ and $S \cap \|\psi\|^t = \varnothing$. Such a set is uniquely decomposed into a non-empty part of $\|\overrightarrow{\varphi}\|^t \setminus \|\overrightarrow{\psi}\|^t$ and a set of semantic values that are neither in $\|\overrightarrow{\varphi}\|^t$ nor $\|\overrightarrow{\psi}\|^t$. There are exactly $2^{2^{card(\textsc{prop})} - card(\|\varphi\|^t \cup \|\psi\|^t)} = 2^{2^{card(\textsc{prop})}} \times \frac{2^{card(\|\varphi\|^t \cap \|\psi\|^t)}}{2^{card(\|\varphi\|^t)+card(\|\psi\|^t)}}$ such sets of semantic values that are in neither interpretations. Hence, there are

$$\frac{2^{2^{card(\textsc{prop})}} 2^{card(\|\varphi\|^t \cap \|\psi\|^t)}}{2^{card(\|\varphi\|^t)+card(\|\psi\|^t)}} \times \left(2^{card(\|\varphi\|^t \setminus \|\psi\|^t)} - 1\right) + \left(2^{card(\|\psi\|^t \setminus \|\varphi\|^t)} - 1\right)$$

continuations admissible for one of the sequence but not for the other (up to semantic equivalence). Distributing $2^{card(\|\varphi\|^t \cap \|\psi\|^t)}$ and using $card(X) = card(X \setminus Y) + card(X \cap Y)$ simplifies the latter expression to

$$2^{2^{card(\textsc{prop})}} \times d_C(\overrightarrow{\varphi}, \overrightarrow{\psi}).$$

Hence, we see that d_C indeed counts, up to semantic equivalence, the number continuations admissible for exactly only one of the two arguments, normalized by the number of sets of semantics values.

In the next section, we will review metrics as shortest path on finite lattices, and see in particular how to generalize d_α, $d_\ominus$ and d_C from set-valued interpretations to more general lattice-valued ones. As a byproduct, we will obtain a general statement of the link between these 3 metrics, as well as a proof that the above continuation metric is indeed a pseudometric (in particular, that it verifies triangular inequality).

4.6.2 Shortest paths in covering graphs

Monjardet (1981) summarizes interesting results concerning metrics on posets and lattices. We will make use of two of these results to shed a different light on the semantic metrics defined in the previous section. In what follows, let $\langle W, \vee, \top, \leq \rangle$ be a bounded semi-lattice with $\top$ as greatest element. For all $x, y \in W$, we say that y covers x iff $x < y$ and $\forall z, x < z \leq y \Rightarrow y = z$. Define also inductively the *rank* of an element, by setting all $<$ -minimal elements of rank 0, and for every y covering a x of rank n, setting y of rank $n+1$. Notice that, given some finite set S and $S' \subseteq S$, $rank(S')$ in the powerset-lattice $\langle 2^S, \subseteq \rangle$, coincide with $card(S')$.

Definition 39 (Covering graph of a semi-lattice)**.** Let $\langle W, \top, \leq \rangle$ be a semi-lattice with $\top$ as greatest element. The *covering graph* $G(W) = \langle V, E \rangle$ of W is such that $V = W$ and $(x, y) \in E$ iff x covers y or y covers x.

Definition 40. An *upper valuation* is an isotone map $v : W \to \mathbb{R}$ such that $\forall z\, z \leq x, y \Rightarrow v(x) + v(y) \geq v(x \vee y) + v(z)$.

Let $G(W) = \langle W, E \rangle$ be the covering graph of W and let v be an upper valuation on W. For each edge $(x, y) \in E$, let the weight function $\omega_v : E \to \mathbb{R}$ induced by v be defined by $\omega(x, y) = |v(x) - v(y)|$. Moreover, let $\pi(x, y)$ be the set of paths from x to y. We make use of two results exposed in Monjardet (1981):

Fact 14 (Monjardet (1981))**.** Let v be an isotone upper valuation. We have:

1. the function $d_v(x, y) = 2v(x \vee y) - v(x) - v(y)$ is positive and verifies the triangle inequality.
2. $d_v(x, y) = \delta_v(x, y) := \min_{p_{xy} \in \pi(x,y)} \sum_{(z_1, z_2) \in p_{xy}} \omega(x, y)$.

Definition 41. If v is an isotone, positive, upper valuation that assigns 0 to minimal elements in the semi-lattice, then the *normalized* distance is defined by

$$d_v^n(x, y) = \begin{cases} \frac{d_v(x,y)}{v(x \vee y)} & \text{if } v(x \vee y) \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Fact 15. If v is an isotone, positive, upper valuation that assigns 0 to minimal elements in the semi-lattice, then the *normalized* distance $d_v^n \geq 0$ and d_v^n verifies the triangle inequality.

This offers a new perspective on the Symmetric difference and Proportional metrics as metrics defined by minimal-weighted paths in the lattice:

Fact 16. Consider the (semi-)lattice $\langle 2^{prop}, \subseteq, \textsc{prop} \rangle$, and the mapping v_0 that assigns to each $V \in 2^{prop}$ its rank. We have

$$d_\ominus(\overrightarrow{\varphi}, \overrightarrow{\psi}) = \delta_{v_0}(\|\overrightarrow{\varphi}\|, \|\overrightarrow{\psi}\|) = card(\|\overrightarrow{\varphi}\| \ominus \|\overrightarrow{\psi}\|)$$

Fact 17. Consider the (semi-)lattice $\langle 2^{prop}, \subseteq, \textsc{prop} \rangle$, and the mapping v_0 that assigns to each $V \in 2^{prop}$ its rank. We have

$$d_\alpha(\overrightarrow{\varphi}, \overrightarrow{\psi}) = \delta_{v_0}^n(\|\overrightarrow{\varphi}\|, \|\overrightarrow{\psi}\|)$$

When the lattice is the lattice of subsets of some set A, the rank of $X \subseteq A$ coincide with its cardinal, hence the two facts above. This suggest d_{v_0} and $d_{v_0}^n$ as natural generalisations of $d_\ominus$ and d_α for more general semi-lattices.

The continuations-based metric is also expressible in these minimal-weighted paths terms, but this requires a little more work:

Fact 18. Let $v : W \to \mathbb{R}$ be an isotone, mapping. The mapping $w : W \to \mathbb{R}$ such that $w : x \mapsto -2^{-v(x)}$ is isotone as well. Moreover, if v is an upper valuation, then so is w.

Corollary 2. Consider the (semi-)lattice $\langle 2^{prop}, \subseteq, \textsc{prop}\rangle$, and the mapping v_0 that assigns to each $V \in 2^{prop}$ its rank, and w_0 defined by $w_0(x) = -2^{-v_0(x)}$. We have

$$d_{\mathcal{C}}(\overrightarrow{\varphi}, \overrightarrow{\psi}) = \delta_{w_0}(\|\overrightarrow{\varphi}\|, \|\overrightarrow{\psi}\|)$$

Corollary 3. $d_{\ominus}, d_{\alpha}$ and $d_{\mathcal{C}}$ are all pseudometrics.

4.6.3 Stronger semantic axioms

We next move to axioms that differentiate between our metrics. We start by considering the following axiom:

$$d(\overrightarrow{\varphi}, \overrightarrow{\varphi}\,\overrightarrow{\psi}) \leq d(\overrightarrow{\varphi}, \overrightarrow{\psi}) \qquad \text{(rebar property)}$$

From the four metrics we have introduced, only d_H^{ham} does not satisfy it.

Fact 19. Let $\|\cdot\|$ be intersective. $d_{\ominus}, d_{\alpha}$ and $d_{\mathcal{C}}$ satisfy (rebar property).

Fact 20. Let $\|\cdot\| = \|\cdot\|^t$. $(\mathsf{L}_{\textsc{prop}}, d_H^{ham})$ does not satisfy (rebar property).

Assume that $\|\cdot\|$ is $\leq$-poset-valued. The following axiom is very mild:

$$\text{If } \|\overrightarrow{\varphi}\| \leq \|\overrightarrow{\psi}\| \quad \text{then} \quad d(\overrightarrow{\varphi}, \overrightarrow{\varphi}\,\overrightarrow{\psi}) \leq d(\overrightarrow{\varphi}, \overrightarrow{\psi}) \qquad \text{(antitonicity)}$$

Fact 21. Let $\|\cdot\|$ satisfy adjunction and right weakening and let d be a semi-pseudometric. If d satisfies (sem$\equiv \Rightarrow$ zero), then d satisfies (antitonicity).

4.6.4 Domain axioms for semantic pseudometrics

Domain axioms require the metric to be well-defined on large portions of the language. Given an interpretation function $\|\cdot\| : \mathsf{L}^* \to D$ of these sequences of sentences into some co-domain, we could expect to have:

$$\text{If } \overrightarrow{\varphi}, \overrightarrow{\psi} \in \mathsf{L}^*, \text{ then } d(\overrightarrow{\varphi}, \overrightarrow{\psi}) \in \mathbb{R}. \qquad \text{(linguistic domain)}$$

It should be realized that the preceding axiom is relatively strong. Consider for example d_H^{ham}.

Fact 22. But $(\mathsf{L}_{\textsc{prop}}, d_H^{ham})$ does not satisfy (linguistic domain).

Fact 23. $d_{\ominus}$, d_{α}, $d_{\mathcal{C}}$ verify (linguistic domain)

Weakening the preceding axiom, without dropping it entirely, can be done if the co-domain of $\|\cdot\|$ is a bounded poset. Recall that $\bot$ denote the least element of a bounded poset.

$$\text{If } \|\overrightarrow{\varphi}\| \neq \bot \text{ and } \|\overrightarrow{\psi}\| \neq \bot, \text{ then } d(\overrightarrow{\varphi}, \overrightarrow{\psi}) \in \mathbb{R}. \qquad \text{(consistent domain)}$$

Fact 24. $(\mathsf{L}_{\textsc{prop}}, d_H^{ham})$ satisfies (consistent domain).

4.6.5 Axioms for set-valued semantic pseudometrics

If $\|\cdot\|$ is set-valued and $target(\|\cdot\|) = \wp(W)$ for some W and (W, δ) is a metric space then we can investigate axioms like the following one considered in Eiter and Mannila (1997):

$$\text{whenever } \|\overrightarrow{\varphi}\| = \{w\} \text{ and } \|\overrightarrow{\psi}\| = \{v\} \text{ then } d(\overrightarrow{\varphi}, \overrightarrow{\psi}) = \delta(w, v) \qquad \text{(EM)}$$

A semantic metric defined as an aggregator of the values of the distance between points in the interpretation of either sequences, will satisfy the preceding axiom.

4.6.6 Signature invariance axioms

The next condition states that the relative proximity of conversations should not depend on irrelevant aspects pertaining to the choice of signature.

$$\text{If } \overrightarrow{\varphi}, \overrightarrow{\psi}, \overrightarrow{\chi} \in \mathsf{L}'^{*} \text{ and } \mathsf{L}' \subseteq \mathsf{L}, \text{ then we have} \qquad \text{(weak sig inv)}$$
$$d_L(\overrightarrow{\varphi}, \overrightarrow{\chi}) \leq d_L(\overrightarrow{\psi}, \overrightarrow{\chi}) \text{ iff } d_{L'}(\overrightarrow{\varphi}, \overrightarrow{\chi}) \leq d_{L'}(\overrightarrow{\psi}, \overrightarrow{\chi})$$

Fact 25. $(\mathsf{L}_{\text{PROP}}(1), \delta_{count})$ and $(\mathsf{L}_{\text{PROP}}(1), \delta_{synt,count})$ satisfy (weak sig inv).

Fact 26. d_H^{ham}, $d_\ominus$, d_α, d_C satisfy (weak sig inv)

4.7 Preservation axioms

4.7.1 Uniform Preservation axioms

The following axiom states that extending conversations with a given piece of information should not change the relative proximity of conversations. Formally:

$$\text{If } d(\overrightarrow{\varphi}, \overrightarrow{\chi}) \leq d(\overrightarrow{\psi}, \overrightarrow{\chi}) \text{ then } d(\overrightarrow{\varphi}\,\overrightarrow{\varphi}_o, \overrightarrow{\chi}\,\overrightarrow{\varphi}_o) \leq d(\overrightarrow{\psi}\,\overrightarrow{\varphi}_o, \overrightarrow{\chi}\,\overrightarrow{\varphi}_o) \qquad \text{(uniform preservation)}$$

But such an axiom can lead to triviality.

Fact 27. Let $\|\cdot\|$ satisfy exchange, contraction and expansion and let d be a pseudometric. If d satisfies (uniform preservation) and (sem$\equiv \Rightarrow$ zero), then whenever $d(\varphi, \chi) \leq d(\psi, \chi)$ then $d(\psi\chi, \varphi\psi\chi) = o$.

The next corollary is slightly technical. Let us introduce a bit of notation. Let $o : \mathsf{L}^* \to \mathbb{R}$ be defined by $o(\overrightarrow{\varphi}) = d(\overrightarrow{\epsilon}, \overrightarrow{\varphi})$, and let $\sim$ be the kernel of o. Let $\leq_o$ be the total pre-order induced by o, with $\overrightarrow{\varphi} \leq_o \overrightarrow{\psi}$ iff $o(\overrightarrow{\varphi}) \leq o(\overrightarrow{\psi})$. Let $[\overrightarrow{\varphi}]$ be the equivalence class of $\overrightarrow{\varphi}$ in $\mathsf{L}^*/\sim$. Let $\downarrow[\overrightarrow{\varphi}] = \{\overrightarrow{\psi} \in \mathsf{L}^* | o(\overrightarrow{\psi}) \leq o(\overrightarrow{\varphi})\}$ and let $\downarrow[\overrightarrow{\varphi}]^*$ be the reflexive transitive closure of $\downarrow[\overrightarrow{\varphi}]$.

Corollary 4. Let $\|\cdot\|$ satisfy exchange, contraction, expansion and $\overrightarrow{\epsilon} - \top$ and let d be a pseudometric. If d satisfies (uniform preservation) and (sem$\equiv \Rightarrow$ zero), then for every $\overrightarrow{\varphi}, \overrightarrow{\chi} \in [\overrightarrow{\varphi}]$, $\overrightarrow{\psi}_1, \overrightarrow{\psi}_2 \in \downarrow[\overrightarrow{\varphi}]^*$ we have $d(\overrightarrow{\varphi}, \overrightarrow{\psi}_1\,\overrightarrow{\chi}\,\overrightarrow{\psi}_2) = o$.

Corollary 5. Let $\|\cdot\|$ satisfy exchange, contraction, expansion and $\overrightarrow{\epsilon} - \top$ and let d be a pseudometric. If d satisfies (uniform preservation), (sem$\equiv \Rightarrow$ zero) and (min sem separation), then for every $\overrightarrow{\varphi}, \overrightarrow{\chi} \in [\overrightarrow{\varphi}]$, $\overrightarrow{\psi}_1, \overrightarrow{\psi}_2 \in \downarrow[\overrightarrow{\varphi}]^*$ we have $\|\overrightarrow{\varphi}\| = \|\overrightarrow{\psi}_1\,\overrightarrow{\chi}\,\overrightarrow{\psi}_2\|$.

Hence (uniform preservation) comes with very disputable consequences. The converse is even more problematic:

$$\text{If } d(\overrightarrow{\varphi}\,\overrightarrow{\varphi}_o, \overrightarrow{\chi}\,\overrightarrow{\varphi}_o) \leq d(\overrightarrow{\psi}\,\overrightarrow{\varphi}_o, \overrightarrow{\chi}\,\overrightarrow{\varphi}_o) \text{ then } d(\overrightarrow{\varphi}, \overrightarrow{\chi}) \leq d(\overrightarrow{\psi}, \overrightarrow{\chi}) \qquad \text{(uniform anti-preservation)}$$

These two axioms are quite demanding.

Fact 28. Let $\|\cdot\| = \|\cdot\|^t$. d_H^{ham} satisfy neither (uniform preservation), nor (uniform anti-preservation).

Fact 29. Let $\|\cdot\| = \|\cdot\|^t$. $d_\alpha, d_\ominus$ and d_C satisfy neither (uniform preservation), nor (uniform anti-preservation).

But the situation is much more radical for (uniform anti-preservation): if the interpretation satisfies very mild conditions: such as contraction, expansion and exchange, then the only semi-pseudometric satisfying (uniform anti-preservation) and (sem$\equiv \Rightarrow$ zero), is the trivial metric.

Fact 30. Let $\|\cdot\|$ be an interpretation satisfying contraction, expansion and exchange and let d be a semi-pseudometric on $target(\|\cdot\|)$. The following are equivalent:

1. d satisfies (uniform anti-preservation) and (sem$\equiv \Rightarrow$ zero)
2. d is the *trivial metric* on $target(\|\cdot\|)$

This result is a very strong argument against the reasonableness of (uniform anti-preservation).

4.7.2 Preservation axioms: close information

The preceding axioms considered extensions of two sequences with the same sequence of formulas. As we have seen, they are too demanding. What if instead, we are interested in the relative effect of extending with a sequences that might be more or less similar to the sequence it is extending. We could expect, that the closer that new sequence is from the original one, the closer the resulting conversation will be from the original one. Or at least that a reverse in respective orderings cannot occur. Formally,

$$\text{If } d(\vec{\varphi},\psi_1) < d(\vec{\varphi},\psi_2) \text{ then } d(\vec{\varphi},\vec{\varphi}\psi_1) \leq d(\vec{\varphi},\vec{\varphi}\psi_2) \qquad \text{(action pref)}$$

Fact 31. d_H^{ham}, $d_\ominus$, d_α, d_C do not satisfy (action pref)

Conversely, we can require the deviation of the resulting sequence to be smaller, whenever the original sequence is closer to the new one by which it is extended.

$$\text{If } d(\vec{\varphi},\chi) < d(\vec{\psi},\chi) \text{ then } d(\vec{\varphi},\vec{\varphi}\chi) < d(\vec{\psi},\vec{\psi}\chi) \qquad \text{(coherent deviation)}$$

Fact 32. Let $\|\cdot\| = \|\cdot\|^t$. d_α satisfy neither (action pref) nor (coherent deviation).

Fact 33. Let d be a semi-pseudometric on $target(\|\cdot\|)$. The following are equivalent:

1. d satisfies (coherent deviation) and the triangle inequality
2. d is the *trivial metric* on $target(\|\cdot\|)$

As we will show, the respective converses of the two preceding axioms are certainly unreasonable.

$$\text{If } d(\vec{\varphi},\vec{\varphi}\psi_1) \leq d(\vec{\varphi},\vec{\varphi}\psi_2) \text{ then } d(\vec{\varphi},\psi_1) \leq d(\vec{\varphi},\psi_2) \qquad \text{(converse strong action pref)}$$

Fact 34. Let d be a pseudometric on $target(\|\cdot\|)$. The following are equivalent:

1. d satisfies (converse strong action pref)
2. d is the *trivial metric* on $target(\|\cdot\|)$

$$\text{If } d(\vec{\varphi},\vec{\varphi}\chi) \leq d(\vec{\psi},\vec{\psi}\chi) \text{ then } d(\vec{\varphi},\chi) \leq d(\vec{\psi},\chi) \qquad \text{(converse coherent deviation)}$$

Fact 35. Let d be a pseudometric on $target(\|\cdot\|)$. The following are equivalent:

1. d satisfies (converse coherent deviation)
2. d is the *trivial metric* on $target(\|\cdot\|)$

4.8 Conjunction and disjunction axioms

4.8.1 Conjunction axioms

Assume that $\|\cdot\|$ is lattice-valued. The following axiom regulates the behavior of the distance with respect to the meet. But as we will see, it is much too demanding.

$$\text{If } \|\vec{\varphi_1}\| \wedge \|\vec{\varphi_2}\| \leq \|\vec{\varphi_1}\| \wedge \|\vec{\varphi_3}\| \quad \text{then} \quad \delta(\vec{\varphi_1}, \vec{\varphi_2}) \geq \delta(\vec{\varphi_1}, \vec{\varphi_3}) \qquad \text{(strong} \wedge \text{rule)}$$

Fact 36. Let $\|\cdot\|$ be lattice-valued and let d be a semi-pseudometric on $target(\|\cdot\|)$. If d satisfies (strong $\wedge$ rule), then whenever $\|\vec{\varphi}\| \leq \|\vec{\psi}\|$, we have $d(\vec{\chi}, \vec{\varphi}) \geq d(\vec{\chi}, \vec{\psi})$ for any $\vec{\chi}$.

Corollary 6. Let $\|\cdot\|$ be lattice-valued and let d be a semi-pseudometric on $target(\|\cdot\|)$. If d satisfies (strong $\wedge$ rule), then whenever $\|\vec{\varphi}\| \leq \|\vec{\psi}\|$, we have $d(\vec{\varphi}, \vec{\psi}) = d(\vec{\psi}, \vec{\varphi}) = 0$

Corollary 7. Let $\|\cdot\|$ be a lattice-valued interpretation satisfying $(\vec{\epsilon} - \top)$ and let d be a semi-pseudometric on $target(\|\cdot\|)$. The following are equivalent:

1. d satisfies (strong $\wedge$ rule) and the triangle inequality
2. d is the *trivial metric* on $target(\|\cdot\|)$

The above facts follow from the equality case in (strong $\wedge$ rule): for any sequences $\vec{\varphi}, \vec{\psi_1}, \vec{\psi_2}$, if $\vec{\psi_1} \wedge \vec{\varphi} = \vec{\psi_2} \wedge \vec{\varphi}$ then $\vec{\psi_1}$ and $\vec{\psi_2}$ have to be equidistant from $\vec{\varphi}$. Removing this assumption yields a weakening of (strong $\wedge$ rule) which no longer support the trivialisation result above:

$$\text{If } \|\vec{\varphi_1}\| \wedge \|\vec{\varphi_2}\| < \|\vec{\varphi_1}\| \wedge \|\vec{\varphi_3}\| \quad \text{then} \quad \delta(\vec{\varphi_1}, \vec{\varphi_2}) \geq \delta(\vec{\varphi_1}, \vec{\varphi_3}) \qquad \text{(weak} \wedge \text{rule)}$$

Fact 37. d_H^{ham}, $d_\ominus$ and d_α do not satisfy (weak $\wedge$ rule).

Fact 38. For any set-valued $\|\cdot\|$, d_C verifies (weak $\wedge$ rule).

Corollary 8. $\|\cdot\|^t$ is an intersective interpretation which yield a d_C that verifies (weak $\wedge$ rule), the triangle inequality and is not trivial.

4.8.2 Disjunction axioms

Assume that $\|\cdot\|$ is lattice-valued. The following axiom is very mild.

$$\text{If } \|\vec{\psi}\| = \|\vec{\varphi_1}\| \vee \|\vec{\varphi_2}\| \text{ then } d(\vec{\varphi}_1, \vec{\psi}) \leq d(\vec{\varphi}_1, \vec{\varphi}_2) \qquad (\vee \text{ rule})$$

Fact 39. Assume that $target(\|\cdot\|) = \wp(W)$ for some non-empty W and that $\|\cdot\|$ is lattice-valued. Let δ be a metric on W. d_H^δ satisfies ($\vee$ rule).

Fact 40. $d_\ominus, d_\alpha$ and d_C also satisfy ($\vee$ rule).

4.9 Future Directions

So far we have focused on isolating an abstract concept of semanticity for metrics. We have explored general axioms that help us express components of this concept. We have also identified more specific axioms that were candidates at defining the contour of a notion of 'good behavior' for semantic metrics, and thus at being criterion for evaluating such metrics. We have done this at an abstract level, considering conversations as sequences of formulas where one conversational agent plays a sequence of formulas after

the other. The conversation thus has the structure of a (syntactic) monoid with a syntactic composition operation of concatenation. Corresponding to sequences of formulas is their abstract interpretation in a different, semantic space; the generic notion of a semantic interpretation furnishes the correspondence, mapping these sequences into the semantic space.

Coming back to the goals outlined in the introduction of this chapter, the next step of our work is to extend this perspective to structures that represent real conversations. We mention a few directions here, each of which can be explored independently. In order to do this, we need to fill in this abstract framework with notions that capture aspects of conversational content at various levels of detail. A first step is to refine the notion of sequence of formulas into something that preserves more of the logical form of conversations. Most models of discourse interpretation assume a more structured representation of conversations, e.g., trees or graphs, in which elementary discourse units are linked together via discourse relations to form more complex discourse units. Using such structures to represent conversations would require us to adapt the structural properties of interpretation functions considered in Section 4.3.2 to be able to reflect the semantics of discourse relations and the units they link together. Second we would need to revisit the axioms that make use of concatenation, replacing the latter with a notion of a graph update or graph extension.

Furthermore, to deal adequately with some natural language phenomena such as questions, commands, agreements and disagreements among speakers, explicit or implicit corrections, it is natural to assume additional structure for semantic spaces, on top of that provided by general lattices. This additional semantic structure could also serve to refine some of our axioms, in particular those making hypotheses on lattice-theoretic relations between the semantic interpretations of two conversations.

Different notions of semantic interpretation carry different amounts of the initial syntactic structure into the semantic space. The classical notion of information content for a discourse erases all structural information, mapping discourses such as *Jane fell because John pushed her* and *John pushed Jane so she fell* into the exact same semantic interpretation (either a set of possible worlds or a set of world assignment pairs as in SDRT and other dynamic semantic theories). Differently structured discourses, even when they share the same meaning, however, may exhibit different semantic and pragmatic behavior, concerning the possibility of future coreferences and of ways to extend the conversation. Intuitions dictate that these features are important for a notion of conversational similarity. It will therefore be important to test metrics defined on more structurally-conservative spaces, for instance conserving some aspects of the conversational graph. These metrics should match intuitions as to how far two real conversations are from each other.

4.10 Conclusions

Our first task was to explore the concept of a semantic metric by identifying a certain number of reasonable axioms that characterize the idea of 'semanticity' for a distance. We clarified the relation between these different axioms and the triangle inequality, and we mapped out a lattice of axioms in terms of their logical strength. Next, we explored a structured list of candidate axioms or desirable properties for any semantic metric. We found several to be too demanding, in the sense that under some structural constraints on the interpretation function and on the distance, they could only be satisfied by the trivial metric. These axioms divide into a certain number of categories. First, we considered a certain number of axioms pertaining to general properties of semantic metrics, including arguably mild assumptions about their structure, their domain and their insensitivity to the choice of signature. Then, we considered preservation axioms that carry a general idea of coherence between the relative proximity of sequences and of their extensions. Finally we considered axioms that are more specific to a lattice- or a set-theoretic approach.

We concentrated on the foundational case of conversations as sequences of propositional formulae

with a classical truth functional interpretation by studying four semantically induced metrics that looked intuitively promising (based respectively on the ideas of symmetric semantic difference, semantic proportionality, Hausdorff metric and on possible continuations). We now have a clear picture of their different behavior. Overall however, these metrics satisfy only few of our axioms that do not lead to a triviality result. One reason for this are the very strong structural hypotheses behind the set-theoretic, classical interpretation of the language of the propositional calculus. A further exploration of these axioms in the context of interpretation into structures like lattices with fewer structural hypotheses and of more general families of metrics remains to be done. We hope that the abstract setting that we have set up in this chapter can serve a first step towards achieving this goal.

4.11 Selected proofs

Proof of Fact 2. Assume that $\forall \vec{\chi}, d(\vec{\varphi}, \vec{\chi}) = d(\vec{\psi}, \vec{\chi})$. In particular $d(\vec{\varphi}, \vec{\psi}) = d(\vec{\psi}, \vec{\psi}) = 0$. Hence by (zero ⇒ sem≡), $\|\vec{\varphi}\| = \|\vec{\psi}\|$. QED

Proof of Fact 3. The right to left direction follows from Fact 2. For the left to right direction, assume that $d(\vec{\varphi}, \vec{\psi}) = 0$ (i). Take any $\vec{\chi}$. By triangle inequality, $d(\vec{\chi}, \vec{\varphi}) \leq d(\vec{\chi}, \vec{\psi}) + d(\vec{\psi}, \vec{\varphi})$. Hence, by (i) we have $d(\vec{\chi}, \vec{\varphi}) \leq d(\vec{\chi}, \vec{\psi})$. Similarly we have $d(\vec{\chi}, \vec{\psi}) \leq d(\vec{\chi}, \vec{\varphi})$. Hence $d(\vec{\chi}, \vec{\psi}) = d(\vec{\chi}, \vec{\varphi})$. But χ was arbitrary, hence for all χ we have $d(\vec{\chi}, \vec{\psi}) = d(\vec{\chi}, \vec{\varphi})$. By (sem separation), it follows that $\|\vec{\varphi}\| = \|\vec{\psi}\|$. QED

Proof of Fact 4. $\|p\neg p\| = \|q\neg q\|$ but $\delta_{count}(p\neg p, q\neg q) = 4$. QED

Proof of Fact 5. $\|p\neg p\| = \|q\neg q\|$ but $\delta_{synt,count}(p\neg p, q\neg q) = 6$. QED

Proof of Fact 6. Take some φ such that $card(\|\varphi\|) \geq 2$. QED

Proof of Fact 7. Assume that $\|\vec{\varphi}\| = \|\vec{\psi}\|$. By (sem preservation) we have $d(\vec{\varphi}, \vec{\chi}) = d(\vec{\psi}, \vec{\chi})$. In particular we have $d(\vec{\varphi}, \vec{\psi}) = d(\vec{\psi}, \vec{\psi}) = 0$ QED

Proof of Fact 8. Assume that $\|\vec{\varphi_1}\| = \|\vec{\psi_1}\|$. Take some $\vec{\chi}$. We have $\|\vec{\chi}\| = \|\vec{\chi}\|$. Hence by (sem induced) we have $d(\vec{\varphi_1}, \vec{\chi}) = d(\vec{\psi_1}, \vec{\chi})$. QED

Proof of Fact 9. Assume that $\|\vec{\varphi_1}\| = \|\vec{\psi_1}\|$ (i) and $\|\vec{\varphi_2}\| = \|\vec{\psi_2}\|$ (ii). By triangle inequality we have:

$$d(\vec{\varphi_1}, \vec{\varphi_2}) \leq \underbrace{d(\vec{\varphi_1}, \vec{\psi_1})}_{0,\ \text{by(sem≡ ⇒ zero)}} + d(\vec{\psi_1}, \vec{\varphi_2})$$

$$d(\vec{\psi_1}, \vec{\varphi_2}) \leq d(\vec{\psi_1}, \vec{\psi_2}) + \underbrace{d(\vec{\psi_2}, \vec{\varphi_2})}_{0,\ \text{by(sem≡ ⇒ zero)}}$$

Hence, $d(\vec{\varphi_1}, \vec{\varphi_2}) \leq d(\vec{\psi_1}, \vec{\psi_2})$. Similarly, we have $d(\vec{\varphi_1}, \vec{\varphi_2}) \geq d(\vec{\psi_1}, \vec{\psi_2})$. QED

Proof of Corollary 1. By Fact 9, d satisfies (sem induced), hence for every $\vec{\varphi_1}, \vec{\varphi_2}, \vec{\psi_1}, \vec{\psi_2}$ with $\|\vec{\varphi_1}\| = \|\vec{\psi_1}\|$ and $\|\vec{\varphi_2}\| = \|\vec{\psi_2}\|$ we have

$$d(\vec{\varphi_1}, \vec{\varphi_2}) = d(\vec{\psi_1}, \vec{\psi_2})$$

It follows that $\dot{d}(\|\vec{\varphi}\|, \|\vec{\psi}\|) := d(\vec{\varphi}, \vec{\psi})$ is well-defined. Moreover for any $\vec{\varphi}$, $d(\|\vec{\varphi}\|, \|\vec{\varphi}\|) = d(\vec{\varphi}, \vec{\varphi}) = 0$. Triangle inequality is proven similarly. QED

Proof of Fact 10. First observe, that by triangle inequality, we have

$$d(\overrightarrow{\chi},\overrightarrow{\varphi}) \leq d(\overrightarrow{\chi},\overrightarrow{\psi}) + d(\overrightarrow{\psi},\overrightarrow{\varphi})$$

Now, assume that $\overrightarrow{\varphi} \equiv \overrightarrow{\psi}$. By (sem$\equiv \Rightarrow$ zero) we have $d(\overrightarrow{\psi},\overrightarrow{\varphi}) = 0$, hence $d(\overrightarrow{\chi},\overrightarrow{\varphi}) \leq d(\overrightarrow{\chi},\overrightarrow{\psi})$. Similarly, $d(\overrightarrow{\chi},\overrightarrow{\psi}) \leq d(\overrightarrow{\chi},\overrightarrow{\varphi})$ which proves (1). QED

Proof of Fact 11. Assume that $\forall \chi$ we have $d(\varphi,\chi) = d(\psi,\chi)$ (i). Take some $\overrightarrow{\chi_1},\overrightarrow{\chi_2}$ with $\|\overrightarrow{\chi_1}\| = \|\overrightarrow{\chi_2}\|$. By (sem$\equiv \Rightarrow$ zero) we have $d(\overrightarrow{\chi_1},\overrightarrow{\chi_2}) = 0$ (ii). By triangle inequality we have:

$$d(\overrightarrow{\varphi},\overrightarrow{\chi_1}) \leq d(\overrightarrow{\varphi},\overrightarrow{\chi_2}) + \underbrace{d(\overrightarrow{\chi_2},\overrightarrow{\chi_1})}_{0,\ \text{by}(ii)}$$
$$d(\overrightarrow{\varphi},\overrightarrow{\chi_2}) \leq d(\overrightarrow{\varphi},\overrightarrow{\chi_1}) + \underbrace{d(\overrightarrow{\chi_1},\overrightarrow{\chi_2})}_{0,\ \text{by}(ii)}$$

Hence $d(\overrightarrow{\varphi},\overrightarrow{\chi_1}) = d(\overrightarrow{\varphi},\overrightarrow{\chi_2})$. Moreover by (i) we have $d(\overrightarrow{\varphi},\overrightarrow{\chi_2}) = d(\overrightarrow{\psi},\overrightarrow{\chi_2})$. Hence $d(\overrightarrow{\varphi},\overrightarrow{\chi_1}) = d(\overrightarrow{\psi},\overrightarrow{\chi_2})$. Since $\overrightarrow{\chi_1},\overrightarrow{\chi_2}$ were arbitrary, it follows by (min sem separation), that $\|\overrightarrow{\varphi}\| = \|\overrightarrow{\psi}\|$. QED

Proof of Fact 18. Let v be an isotone upper valuation and $z \leq x, y$. Assume without loss of generality that $v(y) \geq v(x)$. Let $0 \leq \alpha \leq 1$. Since $2^{-v(y)} \leq 2^{-v(x)}$ and $\frac{1}{\alpha} - 1 \geq 0$, we have

$$2^{-v(y)}(\alpha - 1) + 2^{-v(x)}\left(\frac{1}{\alpha} - 1\right) \geq 2^{-v(y)}\frac{\alpha^2 - 2\alpha + 1}{\alpha} = 2^{-v(y)}\frac{(\alpha+1)^2}{\alpha} \geq 0.$$

Hence,

$$-\alpha \cdot 2^{-v(y)} - \frac{1}{\alpha}2^{-v(x)} \leq -2^{-v(y)} - 2^{-v(x)}.$$

Instantiating this result for $\alpha = 2^{v(z)-v(x)}$ yields after development

$$-2^{v(z)-v(x)-v(y)} - 2^{-v(z)} \leq -2^{-v(y)} - 2^{-v(x)}.$$

Since v is an upper-valuation, we have $-v(x \vee y) \geq v(z) - v(x) - v(y)$ and thus

$$-2^{-v(x\vee y)} - 2^{-v(z)} \leq -2^{v(z)-v(x)-v(y)} - 2^{-v(z)} \leq -2^{-v(y)} - 2^{-v(x)}$$

i.e., $w(x \vee y) + w(z) \leq w(x) + w(y)$ which concludes the proof.
The case $v(x) \geq v(y)$ is symmetrically dealt with. QED

Proof of Corollary 2.

$$\delta_{w_0}(\|\overrightarrow{\varphi}\|,\|\overrightarrow{\psi}\|) = 2w_0(\|\overrightarrow{\varphi}\| \cup \|\overrightarrow{\psi}\|) - w_0(\|\overrightarrow{\varphi}\|) - w_0(\|\overrightarrow{\psi}\|) = -2 \times 2^{-v_0(\|\overrightarrow{\varphi}\|\cup\|\overrightarrow{\psi}\|)} + 2^{v_0(\|\overrightarrow{\varphi}\|)} + 2^{v_0(\|\overrightarrow{\psi}\|)}.$$

Since v_0 coincide with $card()$ on $\langle 2^{prop}, \subseteq, \text{PROP}\rangle$ and $card\|\overrightarrow{\varphi}\| \cup \|\overrightarrow{\psi}\| = card\|\overrightarrow{\varphi}\| + card\|\overrightarrow{\psi}\| - card\|\overrightarrow{\varphi}\| \cap \|\overrightarrow{\psi}\|$, an easy calculation step let us fall back to $d_C(\overrightarrow{\varphi},\overrightarrow{\psi})$. QED

Proof of Fact 20. Let $\overrightarrow{\varphi} := (p_1 \wedge p_2) \vee (\neg p_1 \wedge \neg p_2 \wedge \neg p_3)$ and $\overrightarrow{\psi} := (p_1 \wedge (p_2 \Rightarrow p_3))$, and assume some intersective interpretation of concatenation. We have $\|\overrightarrow{\varphi}\,\overrightarrow{\psi}\| = \|p_1 \wedge p_2 \wedge p_3\|$. $d_H^{ham}(\overrightarrow{\varphi},\overrightarrow{\varphi}\,\overrightarrow{\psi}) = 3$, but $d_H^{ham}(\overrightarrow{\varphi},\overrightarrow{\psi}) = 1$. QED

Proof of Fact 22. For any φ, d_H^{ham} is neither well-defined for $(\varphi,\bot)$ nor for $(\bot,\varphi)$. QED

Proof of Fact 25. Adding a new propositional letter that does not occur in either sequence will not affect the symmetric difference of the range of formulas, nor the symmetric difference of the respective signature. Allowing for the negation of the propositional letter that was previously forbidden will not change the sets either. QED

Proof of Fact 27. Assume that $d(\varphi,\chi) \leq d(\psi,\chi)$ then $d(\varphi\psi\chi,\chi\psi\chi) \leq d(\psi\psi\chi,\chi\psi\chi)$. By exchange, contraction, expansion, and (sem≡ ⇒ zero), we have $d(\psi\psi\chi,\chi\psi\chi) = 0$. Hence $d(\varphi\psi\chi,\chi\psi\chi) = 0$. By exchange, contraction, expansion, and (sem≡ ⇒ zero), we have $d(\varphi\psi\chi,\chi\psi) = 0$. Concluding our proof. QED

Proof of Corollary 4. We only give the idea of the proof. The idea of the proof is to define a linear order on L^* compatible with $\leq_o$. By induction, using Fact (27) we first show the claim for formulas in the same o-equivalence class, then we show that the claim propagate downward, that is for every $\vec{\psi} \in{\downarrow}[\vec{\varphi}]$. Finally we show that the claim propagates with transitive closure. QED

Proof of Corollary 5. Direct from Fact 11 and Corollary 4. QED

Proof of Fact 28. Let $k \geq 2$, $n = 2k$. Now let

$$\varphi := p_1 \Rightarrow (\neg p_2 \wedge \ldots \wedge \neg p_n) \wedge \neg p_1 \Rightarrow (p_2 \wedge \ldots \wedge p_n),$$
$$\psi := p_1 \wedge \neg p_2 \wedge \ldots \wedge p_{2k-1} \wedge \neg p_{2k},$$

$\chi := p_1 \wedge \ldots \wedge p_n$ and $\varphi_0 := p_1$. Since $k \geq 2$ we have

$$1 = d_H^{ham}(\varphi,\chi) < d_H^{ham}(\varphi,\chi) = k, \text{ and,}$$
$$n-1 = d_H^{ham}(\varphi\varphi_0,\chi\varphi_0) > d_H^{ham}(\varphi\varphi_0,\chi\varphi_0) = k = n/2$$

Concluding our proof. QED

Proof of Fact 30. $(1 \Rightarrow 2)$. Take some $\vec{\varphi}, \vec{\psi}$. By contraction, expansion and exchange we have

$$\|\vec{\varphi}\,\vec{\varphi}\,\vec{\psi}\| = \|\vec{\psi}\,\vec{\varphi}\,\vec{\psi}\|$$

Hence by (sem≡ ⇒ zero), $d(\vec{\varphi}\,\vec{\varphi}\,\vec{\psi},\vec{\psi}\,\vec{\varphi}\,\vec{\psi}) = 0$ (i). Hence $d(\vec{\varphi}\,\vec{\varphi}\,\vec{\psi},\vec{\psi}\,\vec{\varphi}\,\vec{\psi}) \leq d(\vec{\psi}\,\vec{\varphi}\,\vec{\psi},\vec{\psi}\,\vec{\varphi}\,\vec{\psi})$ (i). Now, let $\vec{\chi} = \vec{\psi}$ and $\vec{\varphi_0} = \vec{\varphi}\,\vec{\psi}$. By (i) and (uniform anti-preservation) we have

$$\text{If } d(\vec{\varphi}\,\vec{\varphi}\,\vec{\psi},\vec{\psi}\,\vec{\varphi}\,\vec{\psi}) \leq d(\vec{\psi}\,\vec{\varphi}\,\vec{\psi},\vec{\psi}\,\vec{\varphi}\,\vec{\psi}) \text{ then } d(\vec{\varphi},\vec{\psi}) \leq d(\vec{\psi},\vec{\psi}) = 0$$

Hence by (i), $d(\vec{\varphi},\vec{\psi}) = 0$. Concluding the proof for this direction. The other direction is trivial. QED

Proof of Fact 32. Let $n \in \omega$ be such that $n > 5$. Moreover let: $\varphi := (p_3 \wedge \ldots \wedge p_n) \wedge \neg(p_1 \wedge p_2)$, $\psi_1 := (p_1 \vee \ldots \vee p_n) \wedge \neg(p_2 \wedge \ldots \wedge p_n)$ and $\psi_2 := (p_2 \wedge \ldots \wedge p_n)$. We have $d_\alpha(\varphi,\psi_1) = 1 - \frac{2}{n+3} = \frac{n+1}{n+3} > d_\alpha(\varphi,\psi_2) = 1 - \frac{1}{4} = \frac{3}{4}$. But we have $d_\alpha(\varphi,\varphi\psi_1) = 1 - \frac{2}{3} = \frac{1}{3} < d_\alpha(\varphi,\varphi\psi_2) = 1 - \frac{1}{3} = \frac{2}{3}$. QED

Proof of Fact 33. Take some $\vec{\varphi}, \vec{\psi}$. By (coherent deviation) we have

$$\text{If } d(\vec{\varphi},\vec{\epsilon}) < d(\vec{\psi},\vec{\epsilon}) \text{ then } d(\vec{\varphi},\vec{\varphi}\,\vec{\epsilon}) < d(\vec{\psi},\vec{\psi}\,\vec{\epsilon})$$

$$\text{If } d(\vec{\psi},\vec{\epsilon}) < d(\vec{\varphi},\vec{\epsilon}) \text{ then } d(\vec{\psi},\vec{\psi}\,\vec{\epsilon}) < d(\vec{\varphi},\vec{\varphi}\,\vec{\epsilon})$$

Since $\vec{\varphi}, \vec{\psi}$ were arbitrary, it follows that for any $\vec{\varphi}, \vec{\psi}$, $d(\vec{\varphi},\vec{\epsilon}) = d(\vec{\psi},\vec{\epsilon})$. In particular $d(\vec{\varphi},\vec{\epsilon}) = d(\vec{\psi},\vec{\epsilon}) = d(\vec{\epsilon},\vec{\epsilon}) = 0$. Hence by (triangle inequality) we have $\forall\vec{\varphi},\vec{\psi}\, d(\vec{\varphi},\vec{\psi}) \leq d(\vec{\varphi},\vec{\epsilon}) + d(\vec{\epsilon},\vec{\psi}) = 0$. QED

Proof of Fact 34. $(1 \Rightarrow 2)$. Take some $\overrightarrow{\varphi}, \overrightarrow{\psi}$. We have

$$0 = d(\overrightarrow{\varphi}, \overrightarrow{\varphi}\,\overrightarrow{\epsilon}) \leq d(\overrightarrow{\varphi}, \overrightarrow{\varphi}\,\overrightarrow{\psi})$$

Hence by (converse strong action pref) $d(\overrightarrow{\varphi}, \overrightarrow{\epsilon}) \leq d(\overrightarrow{\varphi}, \overrightarrow{\psi})$. But ψ was arbitrary, hence, in particular $d(\overrightarrow{\varphi}, \overrightarrow{\epsilon}) \leq d(\overrightarrow{\varphi}, \overrightarrow{\varphi}) = 0$. But φ was arbitrary as well, hence $\forall \chi d(\overrightarrow{\varphi}, \overrightarrow{\epsilon}) = 0$. Hence by triangle inequality for any formula $\overrightarrow{\varphi}, \overrightarrow{\psi}$ we have $d(\overrightarrow{\varphi}, \overrightarrow{\psi}) \leq d(\overrightarrow{\varphi}, \overrightarrow{\epsilon}) + d(\overrightarrow{\epsilon}, \overrightarrow{\psi}) = 0$. Concluding our proof. QED

Proof of Fact 35. $(1 \Rightarrow 2)$. Take some $\overrightarrow{\varphi}$ and $\overrightarrow{\psi}$. We have

$$0 = d(\overrightarrow{\varphi}, \overrightarrow{\varphi}\,\overrightarrow{\epsilon}) \leq d(\overrightarrow{\epsilon}, \overrightarrow{\epsilon}\,\overrightarrow{\epsilon}) = 0$$

By (converse coherent deviation) $d(\overrightarrow{\varphi}, \overrightarrow{\epsilon}) \leq d(\overrightarrow{\epsilon}, \overrightarrow{\epsilon}) = 0$ (i). Similarly, we have $d(\overrightarrow{\psi}, \overrightarrow{\epsilon}) = 0$ (ii). By (i), (ii) and triangle inequality we have $d(\overrightarrow{\varphi}, \overrightarrow{\psi}) \leq d(\overrightarrow{\varphi}, \overrightarrow{\epsilon}) + d(\overrightarrow{\epsilon}, \overrightarrow{\psi}) = 0$. Concluding our proof. QED

Proof of Fact 36. Take some $\overrightarrow{\chi}$ and assume that $\|\overrightarrow{\varphi}\| \leq \|\overrightarrow{\psi}\|$. Since $target(\|\cdot\|)$ is a lattice. We have $\|\overrightarrow{\chi}\| \wedge \|\overrightarrow{\varphi}\| \leq \|\overrightarrow{\chi}\| \wedge \|\overrightarrow{\psi}\|$. Hence by (strong $\wedge$ rule) $\delta(\overrightarrow{\chi}, \overrightarrow{\varphi}) \geq \delta(\overrightarrow{\chi}, \overrightarrow{\psi})$. QED

Proof of Corollary 6. Assume that $\|\overrightarrow{\varphi}\| \leq \|\overrightarrow{\psi}\|$. Since d satisfies (strong $\wedge$ rule), we have by Fact 36 we have in particular $0 = \delta(\overrightarrow{\varphi}, \overrightarrow{\varphi}) \geq \delta(\overrightarrow{\varphi}, \overrightarrow{\psi}) = \delta(\overrightarrow{\psi}, \overrightarrow{\varphi})$. QED

Proof of Corollary 7. $(1 \Rightarrow 2)$. Take two arbitrary $\overrightarrow{\varphi}, \overrightarrow{\psi}$. By $(\overrightarrow{\epsilon}\text{-}\top)$ we have $\|\overrightarrow{\varphi}\| \leq \|\overrightarrow{\epsilon}\|$ and $\|\overrightarrow{\psi}\| \leq \|\overrightarrow{\epsilon}\|$. Since d satisfies (strong $\wedge$ rule), it follows by corollary 6 that $d(\overrightarrow{\varphi}, \overrightarrow{\epsilon}) = d(\overrightarrow{\epsilon}, \overrightarrow{\psi}) = 0$. By triangle inequality it follows that $d(\overrightarrow{\varphi}, \overrightarrow{\psi}) = 0$. The $(2 \Rightarrow 1)$ direction is trivial. QED

Proof of Fact 38. Take $\overrightarrow{\varphi}, \overrightarrow{\psi_1}, \overrightarrow{\psi_2}$ with $\|\overrightarrow{\varphi}\| \cap \|\overrightarrow{\psi_1}\| \subsetneq \|\overrightarrow{\varphi}\| \cap \|\overrightarrow{\psi_2}\|$. And consider the cardinalities assigned in Figure 4.2. Let $X_1 := \alpha_1 + \eta_0$ and let $X_2 := \alpha_2 + \eta_0$. Using these cardinalities and inserting them in the

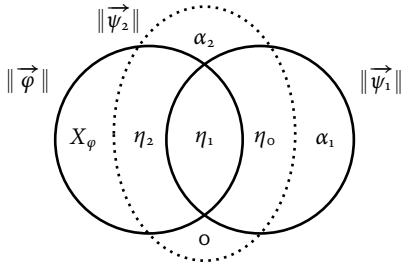


Figure 4.2: Assigning cardinalities to the respective intersections.

expression of the distance, gives us:

$$d_{\mathcal{C}}(\varphi, \psi_1) = \frac{2^{\mathsf{card}(\|\overrightarrow{\varphi}\|)} + 2^{\mathsf{card}(\|\overrightarrow{\psi_1}\|)} - 2 \cdot 2^{\mathsf{card}(\|\overrightarrow{\varphi}\| \cap \|\overrightarrow{\psi_1}\|)}}{2^{\mathsf{card}(\|\overrightarrow{\varphi}\|) + \mathsf{card}(\|\overrightarrow{\psi_1}\|)}}$$
$$= \frac{2^{X_\varphi + \eta_1 + \eta_2} + 2^{X_1 + \eta_1} - 2 \cdot 2^{\eta_1}}{2^{X_\varphi + X_1 + 2\eta_1 + \eta_2}}$$

Similarly:

$$d_{\mathcal{C}}(\varphi, \psi_2) = \frac{2^{X_\varphi + \eta_1 + \eta_2} + 2^{X_2 + \eta_1 + \eta_2} - 2 \cdot 2^{\eta_1 + \eta_2}}{2^{X_\varphi + X_2 + 2\eta_1 + 2\eta_2}}$$

From the two previous expression, after simplifications we find:

$$d_C(\varphi,\psi_1) - d_C(\varphi,\psi_2) = \frac{2^{X_\varphi}\left(2^{X_2+\eta_2} - 2^{X_1}\right) + 2\left(2^{X_1} - 2^{X_2}\right)}{2^{X_\varphi+X_1+X_2+\eta_1+\eta_2}}$$

From the assumption that $\overrightarrow{\varphi} \cap \overrightarrow{\psi_1} \subsetneq \overrightarrow{\varphi} \cap \overrightarrow{\psi_2}$, it follows that $\eta_2 \geq 1$, hence:

$$2^{X_\varphi}\left(2^{X_2+\eta_2} - 2^{X_1}\right) + 2\left(2^{X_1} - 2^{X_2}\right) \geq 2^{X_2+\eta_2} - 2^{X_2+1} + 2^{X_1+1} - 2^{X_1} \geq 0$$

which concludes the proof. QED

Proof of Fact 39. Take $\overrightarrow{\varphi}_1, \overrightarrow{\varphi}_2$. Assume that $\|\overrightarrow{\psi}\| = \|\overrightarrow{\varphi_1}\| \vee \|\overrightarrow{\varphi_2}\|$ (i). By definition, we have

$$d_H^\delta(\overrightarrow{\varphi}_1, \overrightarrow{\varphi}_2) = \max\{\max_{x\in\|\overrightarrow{\varphi_1}\|}\min_{y\in\|\overrightarrow{\varphi_2}\|}\delta(x,y), \max_{y\in\|\overrightarrow{\varphi_2}\|}\min_{x\in\|\overrightarrow{\varphi_1}\|}\delta(x,y)\}$$

Hence, we are in one of two cases.

(1) $\max_{x\in\|\overrightarrow{\varphi_1}\|}\min_{y\in\|\overrightarrow{\varphi_2}\|}\delta(x,y) = d_H^\delta(\overrightarrow{\varphi_1},\overrightarrow{\varphi_2})$, or,

(2) $\max_{y\in\|\overrightarrow{\varphi_2}\|}\min_{x\in\|\overrightarrow{\varphi_1}\|}\delta(x,y) = d_H^\delta(\overrightarrow{\varphi_1},\overrightarrow{\varphi_2})$.

Case 1. There are two subcases.

Subcase 1a. Assume that $d_H^\delta(\overrightarrow{\varphi}_1, \overrightarrow{\psi}) = \max_{x\in\|\overrightarrow{\varphi_1}\|}\min_{y\in\|\overrightarrow{\psi}\|}$ (a). By (i),

$$\max_{x\in\|\overrightarrow{\varphi_1}\|}\min_{y\in\|\overrightarrow{\psi}\|} \leq \max_{x\in\|\overrightarrow{\varphi_1}\|}\min_{y\in\|\overrightarrow{\varphi}_2\|}$$

By (a) and (1) we have $d_H^\delta(\overrightarrow{\varphi}_1, \overrightarrow{\psi}) \leq d_H^\delta(\overrightarrow{\varphi_1},\overrightarrow{\varphi_2})$.

Subcase 1b. $d_H^\delta(\overrightarrow{\varphi_1}, \overrightarrow{\psi}) = \max_{y\in\|\overrightarrow{\psi}\|}\min_{x\in\|\overrightarrow{\varphi_1}\|}\delta(x,y)$ (b). Since δ is a metric we have $\max_{y\in\|\overrightarrow{\varphi_1}\|}\min_{x\in\|\overrightarrow{\varphi_1}\|}\delta(x,y) = 0$. Hence by (i) and (b) we have $d_H^\delta(\overrightarrow{\varphi_1}, \overrightarrow{\psi}) = \max_{y\in\|\overrightarrow{\varphi}_2\|}\min_{x\in\|\overrightarrow{\varphi_1}\|}\delta(x,y)$ (ii). But by (1) and definition of $d_H^\delta(\overrightarrow{\varphi_1},\overrightarrow{\varphi_2})$ we have $\max_{y\in\|\overrightarrow{\varphi}_2\|}\min_{x\in\|\overrightarrow{\varphi_1}\|}\delta(x,y) \leq d_H^\delta(\overrightarrow{\varphi_1},\overrightarrow{\varphi_2})$ (iii). By (ii) and (iii) we have $d_H^\delta(\overrightarrow{\varphi_1}, \overrightarrow{\psi}) \leq d_H^\delta(\overrightarrow{\varphi_1},\overrightarrow{\varphi_2})$.

Case 2. Since δ is a metric we have

$$\max_{y\in\|\overrightarrow{\varphi_1}\|}\min_{x\in\|\overrightarrow{\varphi_1}\|}\delta(x,y) = 0$$

Hence by (2), we have

$$\begin{aligned}\max_{y\in\|\overrightarrow{\psi}\|}\min_{x\in\|\overrightarrow{\varphi_1}\|}\delta(x,y) &= d_H^\delta(\overrightarrow{\varphi_1},\overrightarrow{\varphi_2})\\ &= \max_{y\in\|\overrightarrow{\varphi_2}\|}\min_{x\in\|\overrightarrow{\varphi_1}\|}\delta(x,y)\\ &= d_H^\delta(\overrightarrow{\varphi_1},\overrightarrow{\varphi_2}) \qquad \text{(iv)}\end{aligned}$$

We now consider two subcases.

Subcase 2a. $d_H^\delta(\overrightarrow{\varphi}_1, \overrightarrow{\psi}) = \max_{x\in\|\overrightarrow{\varphi_1}\|}\min_{y\in\|\overrightarrow{\psi}\|}$ (a). But since δ is a metric we have $\max_{x\in\|\overrightarrow{\varphi_1}\|}\min_{y\in\|\overrightarrow{\psi}\|} \leq \max_{x\in\|\overrightarrow{\varphi_1}\|}\min_{y\in\|\overrightarrow{\varphi}_2\|}$ (v). But by definition $d_H^\delta(\overrightarrow{\varphi_1},\overrightarrow{\varphi_2})$ and (2) we have $\max_{x\in\|\overrightarrow{\varphi_1}\|}\min_{y\in\|\overrightarrow{\varphi}_2\|} \leq d_H^\delta(\overrightarrow{\varphi}_1, \overrightarrow{\varphi}_2)$ (vi). By (v), (vi) and (a), it follows that $d_H^\delta(\overrightarrow{\varphi}_1, \overrightarrow{\psi}) \leq d_H^\delta(\overrightarrow{\varphi_1},\overrightarrow{\varphi_2})$.

Subcase 2b. Assume that. $d_H^\delta(\overrightarrow{\varphi_1}, \overrightarrow{\psi}) = \max_{y\in\|\overrightarrow{\psi}\|}\min_{x\in\|\overrightarrow{\varphi_1}\|}\delta(x,y)$ (b). By (iv), it follows that $d_H^\delta(\overrightarrow{\varphi_1}, \overrightarrow{\psi}) = d_H^\delta(\overrightarrow{\varphi_1},\overrightarrow{\varphi_2})$.

Hence in all cases $d_H^\delta(\overrightarrow{\varphi}_1, \overrightarrow{\psi}) \leq d_H^\delta(\overrightarrow{\varphi_1},\overrightarrow{\varphi_2})$. Concluding our proof. QED

Part II

Conversational Games

In chapters 1 to 4 we have been concerned with the logical form of discourse with a focus on its structural properties induced by coherence relations. We have studied the problem of, given two logical representations of the same discourse, determining how they agree and how they differ, and quantifying their similarity or difference.

Chapter 4 made a step in another direction. A discourse is a dynamically evolving object; it has a meaning at each point in time, which evolves as new discourse moves enter the context. We abtracted over a specific account of discourse structure, adopting a perspective of discourses as a sequences of moves equipped with an interpretation function mapping sequences to a given semantic space. Then we investigated the links between different kind of structures for the semantic space (sets, ordered sets, lattices), properties of the interpretation function, for instance regarding the effect of new discourse moves with respect to the structure of the semantic space, and properties of metrics of semantic similarity. We proposed different possible such metrics, the purpose of which is to grasp a scalar notion of 'semantic deviation' between two discourses, or given the abstract perspective that we adopted, two conversations.

Indeed, dialogues and conversations too, are sequences of linguistic moves, with at any point in time, an interpretation. But conversations, in the usual sense of the word, not our abstract perspective, have indeed more structure that what we selectively kept in chapter 4: basically, they involve different speakers, therefore semantic representations must be structured in order to keep tract of who said what. Conversations are also generally thought of has involving a sequence of turns of talk. Their structure is related to that of discourse, as within and accross a speaker's different contributions at different turns of talk, the said speaker generally tries to remain coherent in a sense now familiar from chapter 2. Then again, there is more: for instance, speakers also generally tries to *respond* to the interlocutors moves, in a certain, *relevant* way.

The question we now ask is thus: what is the structure of conversations, and how to formally model the effects, in particular, the semantic effects, of a conversational exchange? We will see that conversational meaning is tied to conversational agents' rationality, and that so-called strategic contexts are a challenge for existing theories. We will propose a new model of conversations that addresses these issues.

Chapter 5

How dialogue differs from texts

Contents

5.1 Specificities of conversations

5.1.1 Agreement, disputes and grounding

Obviously, the fundamental difference between monologues, and conversations of any kind, is the number of participants they involve who actually perform linguistic actions (produce utterances, and respond to others). In the former case, only one agent is conveying content, though there are one or more addressees that interpret what is said. In the second case addressees might speak too. The more striking implication of this, maybe, is the possibility for the listeners of a speaker's contribution to disagree **and** express their disagreement, to which the speaker must then react. In the words of Clark and Brennan (1991): "*When Alan speaks to Barbara, he must do more than merely plan and issue utterances, and she must do more than just listen and understand. They have to coordinate on content*". Hence, while the *informational content* of a monologue is, at least in the correspondence theory of truth, arguably reducible to a single (yet structured and complex) predication, which is, or is not, satisfied by the actual world, the picture is more complicated for conversations.

It takes for a semantic representation to accuretely model the informational content of a conversation, to separate what is commonly agreed on by every agent and what is not. If agents have expressed diverging views, these must be captured and attributed to their resepective owner. This is further complicated by the different level of *grounding* that exist, and concerning each of which agents might signal disagreement or uncertainty: following Clark (1996) there are four levels of grounding:

- of the signal that has been used.
- of the words that this signal convey.
- of the meanings that these words convey.

- of the commitments on the signal's user's part regarding these meanings.

Hence we see that a minimal requirement on a semantic model of conversation, is to separate in the representation the point of view of the conversation's participants. To do this, most approaches, in the vein of Stalnaker (2002), assume a dynamically changing set (with possibly additional structure) of propositions forming the common ground, the set of informations that agents mutually take as presupposed, when planing, performing and interpreting new utterances, and a perspective of conversational dynamics where newly issued content, under specific conditions only, contributes to enrich the common ground, in a specific way.

5.1.2 Illocutionary acts and rationality

A second distinction between monologues and conversations, is the large variety of different illocutionary acts they involve. While such acts are not the prerogative of conversations, as coherence relations, chapter 2 has shown, are a staple of dicourse logical form and typically introduce semantic constraints representing illocutionary effects of the use of a discourse unit in context (*"people mean more than they say"*). Discourse however, does not involve the same diversity of speech acts, such as questions, answers, requests, promises, orders, and others (Searle, 1962, 1965, 1969, *inter alia*) at the same scale as conversation does.

In terms of SDRT (and other models), we will see, this translates as a whole new set of relational speech acts linking one to others contributions, such as relations of question-answer pair (`QAP`), question elaboration (`Q-Elab`), corrections (`Correction`) or acknowledgments (`Acknowledgment`).

This proliferation of illocutionary acts is to be expected given that conversations are really an exchange of *actions*: their effects, rather than solely describing a state of affairs, also change the world as they are performed. These actions are performed for a purpose, which is for their performer to achieve some objectives. As such conversations are thus, essentially, rational activities (Grice, 1975).

Hence, rationality, in return, affects the choice of messages to send and the way agents interpret those messages. A common consequence of these considerations and those formerly made on disagreements and grounding, is that, the semantic effects of a conversational move are not expressible as a standalone content, rather, they must be given in terms of an agent's propositional attitudes toward such a content. The propositional attitudes involved and the way to model changes brought by linguistic actions vary with the proposal; so-called cognitive approaches to meaning ressort to private attitudes forming an agent *cognitive state*, such as beliefs, desires, or intentions (BDI). Gricean, Neo-gricean approaches, and relevance theory (Sperber and Wilson, 1986) fall into this category, but also computational approaches such as (Grosz and Sidner, 1986; Allen and Litman, 1987) which borrow techniques classical in Aritificial Intelligence: BDI, planning and plan-recognition models. Arguably, (see section 6.4), recent game theoretic approaches to conversations (*e.g.,* Benz et al., 2005; Franke et al., 2009) also models meaning in term of belief changes. These models often involve a view of conversations as a joint activity with a common task to solve, or at least a shared interest to attain a given goal. As a result they often take private attitudes as primitive semantic unit, defining locutionary and illucotionary effects in term of agents belief in equilibrium (Franke 2009 however, reintroduces conventional meaning as a *focal point* in the interpretation process). On the other hand Hamblin (1987); Asher and Lascarides (2003); Lascarides and Asher (2009) rather ressort to an objective, public attitude of *commitment*, arguing that conventional meaning remains a primitive which, combined with rationality might form a basis for one to adopt a given belief, rather than the other way around. Correlatively, Asher and Lascarides (2003); Lascarides and Asher (2009) decrease the importance of cooperativity in their account by increasing that of a rich representation of conversations' logical structure.

5.2 Semantic models of conversations

We briefly review two influent models offering fine-grained representations of the semantics and pragmatics of conversations: KoS (Ginzburg, 2012) on the one hand, and D-SDRT Lascarides and Asher (2009) on the other hand.

5.2.1 KoS

KoS is a dynamic semantic model of conversational *interaction*: KoS keeps track of the evolution of a semantic representation of conversational *content*, expressed in the framework of type theory with record (TTR, Cooper, 2005). KoS integrates concepts from a large set of different areas in linguistics, as *e.g.*

- dynamic semantic's view of meaning as update on contexts,
- Clark's considerations on grounding,
- phonology
- syntax (using Head Driven Phrase Structure embedding in TTR)
- Pragmatics' common ground.
- An account of discourse structure using a Questions Under Discussion framework (QUD, Roberts, 2012)

Indeed, one of KoS major contributions, is to achieve a simultaneous and coherent application of such various sources of linguistic knowledge to the modeling of challenging semantic phenomena such as clarification questions, which can target every level of semantic interpretation. The very large scope of KoS representations, which include descriptions of signals at every level: phonems, syntactic units, semantic representation, illocutionary effects and metatalk relations, allows the theory to achieve an unprecedented explicative power of semantics in spoken dialogs in their full complexity, *i.e.* including, and accounting for, hesitations, disfluencies or repairs, among other such phenomena.

KoS implements the dynamics of conversations through a *grammar of interaction*. Agents are assigned a Dialog Game Board (DGB) that dialog moves affect in a specific way, modeled in the theory through a pair of preconditions and effects. Interestingly, DGB involve a public component, and a private one. Two agents' public component need not match perfectly, as a consequence, though correlated in some way, agents' public attitudes need not involve only facts of the *common ground* as in more classical approaches. This is, we think, an attractive feature of the model, and one that the commitment-based semantics we propose in part III will reproduce, in a model-theoretic setting.

The private and public parts of the DGBs together represent a full informational state for an agent. In that sense as well, the theory is one of interaction: assuming two (or more) agents represented with their complete informational states, the theory describes how conversational interaction will make these informational states dynamically evolve.

In contrast, the D-SDRT approach that we present next, is more restricted in scope, but focus on an objective notion of meaning that generally do not require knowledge of the private component of agents' informational states.

5.2.2 D-SDRT

D-SDRT augment SDRT that we presented in chapter section 2.3 with an account of conversations, the semantics of which is expressed in terms of commitments.

Semantic representation in D-SDRT are called D-SDRSs; the switch from SDRS to D-SDRSs involve two changes:

- At every point in the conversation, each conversational agent is assigned her own SDRS.
- The interpretation of D-SDRS is now expressed in terms of context-change potential over p-tuple of (world, assignment) pairs with p the number of agents involved in the conversation.

More formally, recall our definition of an SDRS: with Π be an infinite countable set of discourse labels, and $\mathcal{R}$ be a set of rhetorical relations, An SDRS is a tuple $S = \langle A, O, E, \pi_{\text{top}}, f \rangle$, with (roughly) A a set of labels, $\pi_{\text{top}} \in A$, O a set of edges forming a tree over A, E a set of directed edges forming a directed acyclic graph over A, and f assigning content to EDUs (plus other additional constraints, see section 2.3). D-SDRSs are defined over SDRSs as follows:

Definition 42 (Syntax of D-SDRSs)**.** Assume a set of agents I. A D-SDRS is a tuple $\langle n, \text{spk}, T, A, O, E, \pi_{\text{top}}, f \rangle$ where

- spk assigns each turn $k \in [0, n]$ a unique speaker $i \in I$.
- $S = \langle A, O, E, \pi_{\text{top}}, f \rangle$ is an SDRS
- T is a mapping from $[1; n] \times I$ to subsets of A that assigns a pair of (turn, speaker) (k, i) a subset $A_{k,i}$ such that $S_{\restriction A_{k,i}}$ is an SDRS. We let $S_{k,i} = S_{\restriction A_{k,i}}$.
- With $<_S$ denoting 'spoken order', *i.e.* the order of production of EDUs by speaker, and $\text{last}(X)$ the last EDU label in the set $X \subseteq A$ w.r.t. to $<_s$, we have $\text{last}(A_{n,\text{spk}(n)}) = \text{last}(A)$, and $\text{last}(A_{k,i}) < \text{last}(A_{k+1,\text{spk}(k+1)})$ for every $k < n$ and $i \in I$.

Crucially the syntax of D-SDRS allow conversational agents to share labels across their SDRSs, and their respective SDRSs, at each turn must correspond to a fragment of a common, complete SDRS. Intuitively, this corresponds to a classical, binary view on grounding implemented through discourse labels: at some turn, either every agent have integrated a label into their respective SDRS, in which case we might say that the label is *grounded*, in the sense that discourse participant at least agree that this label and content are part of the context. Note that if a label is *grounded* in that sense then so are all the relations that this label bear to other *grounded* constituents, or at least one of them fails to acknowledge (at this point) that the speech act that the label represents occured.

D-SDRSs interpretation is expressed as a context change potential over *commitment slates*, where the change of each entry in the commitment slates relies on the interpretation of SDRSs in section 2.3:

Definition 43 (Interpretation of D-SDRSs)**.** Let $p = card(I)$. A *commitment slate* is a p-tuple of world-assignment pairs (w, f). Given a commitment slate σ Let $\sigma[p]$ denote the kth component of σ.

Let $D = \langle n, \text{spk}, T, A, O, E, \pi_{\text{top}} \rangle$ be a D-SDRS. The context change potential of D is provided by:

$$\sigma \; [\![D]\!]_d \; \sigma' \text{ iff } \forall i \in I \; \sigma[i] \; [\![A_{n,i}]\!] \; \sigma'[i]$$

With $[\![\cdot]\!]_d$ the interpretation of D-SDRSs and $[\![\cdot]\!]$ the interpretation of (regular) SDRSs.

Intuitivively, a commitment slate σ represent a set of commitments for each agent: agent i is commmited to the set of propositions represented by the DRSs that the (world, assignment) pair $\sigma[i]$ validates. Moreover, the content change potential at turn k is given by that of the D-SDRS $\langle k, \text{spk}, T, S_{\restriction A_{k,\text{spk}(k)}} \rangle$, which gives an idea of the dynamic evolution of agents' commitments throughout the conversation.

What we have called a binary view on grounding is tied to the abscence of a commitment modality into the logical language: commitments come as an external machinery provided by commitment slate, and are

not objects that agents might talk about (since context change potential is always evaluated always entries of commitment slates **separately**). In part III of this thesis, we will come back to these considerations and propose a way to add ambiguity and commitments about commitments to the picture.

Chapter 6

Implicatures, games and strategic contexts

Contents

6.1 Game theoretic pragmatics

Assuming that conversationalists are rational, what they say and how they interpret what is said should follow as actions that maximize their interests given what they believe. Conversational moves should be calculated via an estimation of best return given what other participants say, which is a natural setting for game theoretic analyses.

Game theory has had several applications in pragmatics Lewis (1969); Parikh (1991, 2000, 2001); Benz et al. (2005); Franke (2008); Franke et al. (2009); van Rooij (2003, 2004). Much of this literature uses the notion of a signaling game, which is a sequential (dynamic) game in which one player with a knowledge of the actual state sends a signal and the other player who has no knowledge of the state chooses an action, usually an interpretation of the signal. The standard set up supposes that both players have common knowledge of each other's preference profiles as well as their own over a set of commonly known set of possible states, actions and signals. The economics literature contains a detailed examination of signaling games, Spence (1973); Crawford and Sobel (1982); Farrell (1993a); Rabin (1990), to name just a few important papers in this area.

6.2 Strategic contexts

Conversations often involve an element of planning and calculation of how best one can achieve one's interests. Previous sections have mentioned game-theoretic approaches to the semantics and pragmatics of conversations which represents the very meaning of conversational moves as a byproduct of contextual strategic reasoning. However, such models implicitely assume that conversationalists share a common interest in solving a given task, an assumption which echoes the gricean maxim of cooperativity. How these models behave when cooperativity is dropped, is something that we will discuss a little more ahead

in this chapter. We only remark for the time being, that they have not been purposedly constructed or extensively tested for such kind of non-cooperative data.

More generally, while there is a large literature in linguistics and in AI on cooperative conversation stemming from Grice (1975), there is little theoretical and formal analysis of conversation in non-cooperative situations. The work of Traum and Allen (1994), where cooperativity is determined only by the social conventions guiding conversation, obligations that do not presuppose speakers to adopt each other's goals, is an important exception. But still, the formal structure of such conversations remains largely unexplored. From this section onwards, we will therefore be interested in how conversations proceed in a setting where, precisely, a general assumption of cooperativity does not hold, *i.e.* in which dialogue agents cannot assume that their interests coincide with those of their interlocutors. We will argue that a form of *communication* survives in those context, at least in a sense that we will define, and propose a formal theory of message exchange in settings where agents do not necessarily share interests and goals. We think that the study of such settings constitutes a promising starting point for achieving a general model of conversation.

In particular, a little explored element in linguistics is the general "shape" of a conversation, its overall structure and the effects of this structure on content. The goals of conversational participants and the context of moves they have already made explain why they make the subsequent discourse moves they do and give a coherence to the conversation as a whole. For conversations where agents share conversational goals and interests, a broadly Gricean answer explored by Grosz and Sidner (1986); Grosz and Kraus (1993) *inter alia* is that the discourse is organized around a problem that it is in the common interest of the participants to resolve; the structure of the conversation reflects the structure of the decision problem, or rather the reasoning of conversational participants to construct a plan that solves the common decision problem.

Non cooperative conversations, conversations where cooperativity or shared goals and interests cannot be assumed, don't instantiate reasoning about a common decision problem. Consider a debate between two political candidates. Each candidate has a certain number of points she wants to convey to the audience; each wants to promote her own position at the expense of the other's. Strategic conversations are also reactive: to achieve their goals, each participant needs to plan for anticipated responses from the other. To explain "what is going on" in such a conversation, we need to appeal to the participants' discourse goals, which may depend on the goals of the participants' interlocutors. Similar strategic reasoning about what one says is a staple of board room or faculty meetings, bargaining sessions, and even conversations with one's children. These observations indicate that strategic conversations are games, and debates are typically o sum games. Typically only one agent can win, though there may also be draws. Such conversations are common.

Grasping the general conversational goals of conversationalists does not suffice, however, to determine the structure of a conversation. Since conversations should be the result of rational inference to the best means for achieving one's conversational goals given one's information about the discourse context, particular *linguistic moves* in a conversation should be related to an overall conversational goal. For cooperative conversations, we need to describe the linguistic reflection of the reasoning about a common decision problem, and this means we need to talk about the way clauses in a text rhetorically relate to each other and how such related clauses can combine to form more complex discourse units bearing rhetorical relations to other discourse constituents in a way that has become familiar not only from Grosz and Sidner but from theories of discourse structure like the ones we introduced in chapter 2. The interaction between goals and particular moves is important for understanding monologue as well, as one can ask what "problem" the author was trying to solve in a particular passage; there is a close correspondence between a coherent text's discourse structure and the text's "goal". We aim to tell a similar story for conversations in not necessarily cooperative settings. Given certain general conversational goals for our conversational participants, we want to track how particular discourse moves detailed in a theory like SDRT takes one dialogue agent towards her conversational goals or thwarts them.

To get a better idea of the structure of conversations in strategic settings, we start from two intuitions. First, there is a strong intuition that many strategic conversations have a determinate outcome. One dialogue agent can "win" if she can play certain conversational moves; and if she doesn't, she loses. Second, another equally strong intuition is that, in many conversations, some conversational strategies, and some winning conditions or conversational goals, are more complex and more difficult to achieve than others. Understanding and categorizing such winning conditions and their strategies are an important part of understanding the large scale structure of conversations. In addition, they also determine whether a conversational agent has won the strategic conversational game, which is one important communicative effect of the conversation. But how can we measure or compare such strategies? We will enforce to systematize these intuitions and offer an answer to this question.

To this aim, we first introduce a set of key examples of strategic conversations, as well as their intuitive winning conditions. This will help us fix intuitions and we will regularly come back to and discuss these examples in subsequent sections and chapters.

6.3 Key examples and intuitions

6.3.1 examples

We present here examples of strategic conversations of two kinds: linguistic reports of real or constructed conversations and their context, on the one hand, and unrealized strategic contexts and there winning conditions (more like thought-experiments) on the other hand. Our first and simplest example is of the latter kind:

(6.3.1) Suppose a candidate, Candidate A, has a joint interview with another competing candidate, Candidate B, for an academic position. Suppose Candidate A has proved an important theorem and she knows that during the interview if she can mention this, she will have "won" the interview by getting the job over the smarter candidate B, as long as she can mention this fact, no matter at what point of the meeting she says so. This is her winning condition.

In the following example from Solan and Tiersma (2005), a prosecutor wants Bronston to say whether he had a bank account in Switzerland or not; Bronston does not want to make such an admission. His winning condition is to not answer the question directly, but only to implicate an answer that he doesn't have a bank account. He does not want to commit either to having or to not having a bank account.

(6.3.2)a. Prosecutor: Do you have any bank accounts in Swiss banks, Mr. Bronston?
b. Bronston: No, sir.
c. Prosecutor: Have you ever?
d. Bronston: The company had an account there for about six months, in Zurich.

A non-courtroom variant of example (6.3.2) is example (6.3.3). The background is that Janet and Justin are a couple, Justin is the jealous type, and Valentino is Janet's former boyfriend.[23]

(6.3.3)a. Justin: Have you been seeing Valentino this past week?
b. Janet: Valentino has mononucleosis.

Janet's response implicates that she hasn't seen Valentino, whereas in fact though Valentino has mononucleosis she has been seeing him.

[23] Thanks to Chris Potts and Matthew Stone for this example.

Consider a *voire dire* examination in a medical malpractice suit where the plaintiff lawyer (LP) has as a goal to return repeatedly to the topic about the division of a nerve during a surgery. This goal has a further objective of getting the witness (D) to characterize the surgical operation as incompetent and mishandled. Repeatedly coming back to the topic can wear D down as actually happened in the case we cite.

(6.3.4)a. LP: And also, he put an electrical signal on that nerve, and it was dead. It didn't do anything down in the hand, it didn't make the hand twitch?
b. D: Correct.
c. LP: And we know in addition to that, that Dr. Tzeng tore apart this medial antebrachial cutaneous nerve?
d. D: Correct.
e. LD: Objection.
f. THE COURT: Overruled.
g. D: Correct. There was a division of that nerve. I'm not sure I would say "tore apart" would be the word that I would use.
h. LP: Oh, there you go. You're getting a hint from your lawyer over here, so do you want to retract what you're saying?

The defendant was resisting this line of attack relatively well, but then made an error by agreeing to LP's loaded question.

During the Dan Quayle-Lloyd Bentsen Vice-Presidential debate of 1988, Quayle was repeatedly questioned about his experience and his qualifications to be President. Quayle attempted to compare his experience to the young John Kennedy's to answer these questions; his winning condition was probably to suggest with this comparison that like Kennedy he was a worthy Presidential candidate. Part of his goal too was to have this comparison pass without criticism (perhaps because he couldn't defend it adequately), and so it was indirect. However, Bentsen made a discourse move that Quayle didn't anticipate.

(6.3.5)a. Quayle: [. . .]the question you're asking is, "What kind of qualifications does Dan Quayle have to be president," [. . .] I have far more experience than many others that sought the office of vice president of this country. I have as much experience in the Congress as Jack Kennedy did when he sought the presidency.[. . .]
b. Bensten: Senator, I served with Jack Kennedy. I knew Jack Kennedy. Jack Kennedy was a friend of mine. Senator, you're no Jack Kennedy.
c. Quayle: That was really uncalled for, Senator.
d. Bentsen: You are the one that was making the comparison, Senator — and I'm one who knew him well.[. . .]

Bentsen's surprise move successfully attacked Quayle's strategy to establish a comparison between himself and John Kennedy. Quayle had no effective defense and lost the debate handily.

(6.3.6)Allegedly, the physicist and Nobel laureate Richard Feynmann decided the topics of his next lecture in advance and prepared for it for over 8 hours. However, when he entered the class he would start off with: "So what shall we discuss today?" But he would always have a strategy to steer the conversation to the topics he had prepared for, whatever his students, who always wanted to stump him (and so had opposing interests to Feynmann's), would answer. Feynmann's winning condition was eventually to get to his prepared topic and stick to it for the remainder of the lecture.

Our examples so far have described or been excerpted from finite conversations that are relatively circumscribed. But conversations can occur over a much longer period, say over an entire Presidential campaign as in our next example. Nevertheless, they are still *linguistic conversations*.

(6.3.7) Recall President Clinton's adage "it's the economy stupid." What Clinton meant is that he should keep the conversation focussed on questions concerning the economy in the extended debate between his Democratic team and the opposing Republican one during the 1992 Presidential campaign. As long as Clinton was able to bring the debate repeatedly back to a discussion of the economy, he achieved his winning condition.

From the examples listed above, we draw a handful of principles that we think are essential constitutive properties of strategic settings and that a conversational model needs to take into account:

6.3.2 Important features of conversations

We claim the following are important features of strategic conversations (and perhaps of conversations generally).

(I) People have conversations for purposes. Their conversations are successful when they achieve those objectives. Crucially, some of these objectives involve commitments to contents by other conversational participants. In all conversations, including those where one person's gain from the conversation is another person's loss, the interlocutors' contributions force them to commit to certain contents, which are the conventional meanings and implicatures of their utterances in the context.

(II) In principle, conversational players have no limits on the length of their intervention, though they are finite. In practice exogenous time limits may be imposed.

(III) Players can in principle "say anything" during their conversational turn, though what they say may very well affect whether their conversation is successful or not.

(IV) While conversations are finite, they may have no designated "last turns;" conversational agents cannot in general foresee who will "have the last word." Hence, people strategize in conversations even when they can't anticipate when the conversation will end, what possible states might arise, or what utterances their opponent will consider.

How well then do existing models, especially those based on signalling games, fare with respect to these principles and strategic settings? The next section discuss this question.

6.4 Difficulties for signaling-based accounts

Signaling games and related model that we discussed in section 6.4 implements both the choice of a given dialogue moves and the computation of its meaning, as the maximization of rational agents' interests in a given context. This context is formally expressed as a game arena, set of possible moves and payoffs for the agents. As such, signaling games have proved valuable for many issues, particularly the formalization of an effective computation of various kind of implicatures. Despite all this, we will see that they do not offer a straightforward way to encode the principles that we outlined in the previous section, regarding strategic contexts. As a consequence, we will propose in this thesis a models that differs from signaling games in many aspects. Our proposed model won't contradict the predictions of signaling models, however, but rather propose a natural and convenient way of addressing situations that are not transparently expressible as a signaling game's context. We now explain why the strategic contexts we consider fall into this category.

A game requires, in order to be a reasonable candidate for modeling non-cooperative contexts, that its structure encodes the players' divergent preferences. As emphasized earlier, the most intuitive way of doing that is to assume a 0 sum game. Signaling games however predict that no communication happens in

such games: it can be shown that in equilibrium[24] the sending of any message has no effect on the receiver decision.

An immediate corollary is: assuming that the sender has the possibility of (costlessly) not sending a message and that the sending of any message has at least an infinitesimal cost, ϵ, makes it optimal for the sender to not send anything. This leads to obvious, unintuitive and irrational consequences. Hence, the most straightforward way of setting up non-cooperativity makes communication of any kind is impossible in a signaling game. This means that non-cooperativity of the sort we are interested here should not translate as 0-sum utilities in a signaling model.

Still, there is between perfectly aligned utilities and 0-sum games, a space of games with partially aligned utilities which could encode (some) lack of cooperativity into the context while still allowing for communication to take place. Notice that, crucially, yielding the right equilibrium is not the only demand to put on the game structure: a precise justification of the chosen utility profile is needed, just as much. In order to use games as part of a general theory of meaning, one has to make clear how to construct, for each situation of interest, an adequate game-context. Importantly, this includes providing an interpretation of the game's ingredients (types and actions) and explaining why the utility profiles fits the situation to be modeled. Franke (2009), for instance, associates in a principled way an *interpretation game* to a given utterance. Interpretation games form a subclass of signaling models assuming a specific class of sender types actions and preferences. They intend to encode a "canonical context" for an utterance, in which relevant conversational implicatures may be drawn. In interpretation games, the full game structure is determined by the set of sender types: there is a bijection between the set of receiver actions and the set of sender types, and the utility profile is such that both the receiver and sender get rewarded if they coordinate on the sender actual type, and do not gain anything otherwise.

Such a setting is very intuitive and interestingly does not seem to require further precision on what exactly it means for the receiver to take the action a_t associated with receiver type t and why such an action should indeed maximize the receiver payoff if t is the sender's actual type. Of course, one can still wonder whether performing a_t means, for instance, that the receiver believes that the actual state is t (let us call this option 1), or that the receiver interprets the sender's message as a commitment to the actual state being t (option 2), or even that she herself commits to the sender commiting to the actual state being t (option 3). But, despite this ambiguity between, in principle, distinct ways of understanding action a_t, there is, in interpretation games, no necessity to settle the question, because, for **Gricean agents**, the different options collapse. A Gricean sender should intend to commit to what he believes is true (sincerity), cooperativity should make a Gricean receiver intend to interpret the sender commitment as what the sender intends to communicate, and belief in the sender's sincerity should make him believe that the sender believes in what he has committed to. Hence options 1 and 2 collapse. Finally the receiver's sincerity and cooperativity ensure that (at least if the sender's needs her to), she acknowledges that the senders' commits to what she interprets the sender to commit to (namely, that t is the actual sender's state). Hence option 2 implies option 3, and the other way around follows from the receiver's sincerity again. Therefore, the games structure as it stands seems to offer a perfectly adequate level of abstraction.

But things become much more intricate as soon as one is considering potentially non-Gricean players, and this makes the task of understanding and providing justification for a (partially) unaligned utility profile much more involved. It depends on what one takes actions and types to represent. Recall example (6.3.2) and imagine, for the sake of argument, that we want to model Bronston's answer with a signaling game involving two sender types: t_{bank} and $t_{\neg\mathsf{bank}}$, two corresponding interpretative (or acknowledging) actions a_{bank} and $a_{\neg\mathsf{bank}}$, and three possible messages, m_{yes}, m_{no} and m_{company}. These messages are respectively

[24] Assuming bounded rationality of conversational agents may restore an effect to messages: for instance the Iterative Best-Response model in Franke (2009) allows a level 2 sender to misdirect a less sophisticated level 1 receiver. However, we are convinced that the conversational examples presented in this article are compatible with a common belief in rationality and require an analysis making such an assumption.

true in the sets of states $\{t_{\mathsf{bank}}\}$, $\{t_{\neg\mathsf{bank}}\}$ and $\{t_{\mathsf{bank}}, t_{\neg\mathsf{bank}}\}$. Assume also that we want to accommodate a fear of perjury on Bronston's part into the game context. Consider first that performing action $a_{\neg\mathsf{bank}}$ means for the receiver to update his belief to includes that $t_{\neg\mathsf{bank}}$ is the actual sender's type, or at least, to subsequently acts as if it were the case. Under such an interpretation, if the sender, Bronston, sends m_{no} and the prosecutor takes in return action $a_{\neg\mathsf{bank}}$, should Bronston fear being charged with perjury? Intuitively no, because such an attack would indicate an inconsistent belief of the prosecutor that $t_{\neg\mathsf{bank}}$ holds (because the action he took is interpreted as such) and does not hold at the same time.[25] Then again, if actions are to be interpreted at the level of public commitments (say, using option 3 of the previous paragraph), taking action $a_{\neg\mathsf{bank}}$ after receiving m_{no} commits the prosecutor to the proposition that Bronston is committed to $t_{\neg\mathsf{bank}}$. This does not imply that the prosecutor believes (or commits to) the latter state to be actual. Hence, if the prosecutor takes this action, he is susceptible to attack Bronston for perjury by committing to Bronston's actual type being in fact t_{bank} and not $t_{\neg\mathsf{bank}}$. Bronston's payoff in that case should depend on whether the prosecutor will indeed charge him with perjury and how bad the consequences will be. If the prosecutor has solid arguments to show that t_{bank} is the actual state, then Bronston should fear such a continuation and have a very low payoff for the triple $\langle t_{\mathsf{bank}}, m_{\mathsf{no}}, a_{\neg\mathsf{bank}}\rangle$. Bronston should be better off with the triple $\langle t_{\mathsf{bank}}, m_{\mathsf{company}}, a_{\neg\mathsf{bank}}\rangle$ as, in that case, he disposes of a way to defend himself against a charge of perjury, which consists in commiting that he never committed to $t_{\neg\mathsf{bank}}$. Put another way, he can argue that the prosecutor with $a_{\neg\mathsf{bank}}$, committed to something false –namely that Bronston committed to $t_{\neg\mathsf{bank}}$. Notice that such a defense is not a very good option when Bronston sends m_{no}, because in that case, it requires Bronston to say something false[26]. Notice also, that ressorting to this defense should bear a non-neglictible cost: allthough he avoids perjury, Bronston can be asked to answer why he did not respond to the the prosecutor question in the first place and/or why he did not immediately correct the prosecutor after the latter performed $a_{\neg\mathsf{bank}}$.

These considerations illustrate two things: first, if not all agents conform to Gricean maxims, different choices in the way to interpret actions and types yield different games context with different predictions. Hence it becomes primordial to make precise what the exact set of actions is and what they represent —something which may vary according to the nature of the player's objectives (commitments, beliefs, both, something else, …). Second, the payoffs of the sender and receiver may depend on subsequent actions, which requires that the possible outcomes of the signaling games encode all possible relevant continuations of the conversation. None of this is self-evident (and we will examine some reason for why it is so) and makes a systematic construction of a game context much more difficult than in the cooperative case. These difficulties echoes the close correspondence between a general formalization of Gricean principles and that of games with shared interests that Asher and Lascarides (2012) establishes. This doesn't entail that in strategic settings, Gricean principles don't ever apply, but the result does establish that one shouldn't count on Gricean principles as operative; in general one can't assume that players are maximizing quality, quantity or relevance (to one's own conversational ends).

Furthermore, there is a crucial difference between being a non-Gricean speaker, and admitting to being so. A player's conversational objectives are very likely to include not making such an admission. Conversations can thus involve sorts of hide-and-seek games where agents try to expose the "bad" behavior of their opponent while making themselves look good. In other words, bad (typically, non-Gricean) behavior is licenced as long as it remains hidden or deniable. In example (6.3.2), if the prosecutor rejects Bronston's indirect answer, he commits at the same time that Bronston's cooperativity is, at least, subject to caution. If Bronston then admit having had a swiss bank account, he justifies this cautious attitude, and commits to

[25] We assume here that the prosecutor has an interest to charge Bronston with perjury only if he believes that Bronston actually performed perjury. One can relax this assumption, but that would mean that the prosecutor's beliefs are irrelevant to his subsequent moves and that the commitments-related interpretation of actions should be considered here.

[26] and following the logical model of commitment that one adopts, it can even make him inconsistant. See part III of the present thesis.

the proposition that he was not cooperative in giving the indirect answer. In such a context, the prosecutor intuitively should claim that Bronston is being non-cooperative but Bronston should try to avoid admitting that he is. In analogous contexts occurring outside of the courtroom, it might be rational for an interrogator who cares for his reputation or his interlocutor's friendship to prefer a misleading answer over formulating a public accusation of non cooperativity that he cannot prove.

Signaling games being one-shot games basically leaves the modeler with two choices: either the game's payoffs are locally determined, and then the game, by choice of design, does not evaluate the long term impact of a triple ⟨type,message,action⟩, $\langle t, m, a\rangle$, or the payoffs are global, but then must somehow be computed taking every possible continuation of $\langle m, a\rangle$ into account (if this is possible at all, which is discussed more at length in section 7.2.1). There is also an afferent problematic asymmetry of the sender and receiver in such games. While, in signaling games, the sender, sending a message, reveals some information about his type and associated preference profile that the receiver may seize on, the latter does not reveal anything that is not already common belief when he chooses an action. There is anyway no subsequent move of the sender to reward or punish such a revelation by the receiver.

Again, this is problematic because it blurs the frontier between public and private information which an accurate analysis of strategic settings requires to be clearly marked. While types, preferences and beliefs are intuitively private, messages are intuitively public. The nature of receiver's actions however, as we have seen, is not clear: belief update or interpretative actions are private actions whereas acknowledgments, responses are not. Moreover neither option is plainly satisfactory (for a full account of strategic context): if actions are private updates, then they are not sufficient on their own to constrain subsequent conversational moves, unless the model is supplemented with a theoretical link between private and public attitudes (once again, this is precisely what a formalization of the Gricean maxims provides **for cooperative settings only**). If actions are public, then the rational receiver taking a particular action reveals, at the same time, having a certain belief about the sender's state, namely one that makes the choosen action rational. Even if we assume this belief to be common to the sender and the receiver, it does not mean that either of the two is comfortable with making this belief into a public commitment. There can be numerous reasons for avoiding such an admission: for instance, a third party monitoring the conversation, and/or politeness constraints preventing the players to put their cards on the table. Imagine two agents holding each other in very low regard, with a common belief thereof: it seems reasonable to imagine keeping the conversation polite and cordial as one of their objective, restricting without voiding the moves they might use toward a second objective of defending a position at the expense of the other. They should, in particular never explicitely commit to what they think of each other despite being fully aware of it on both sides. Yet, due to signaling games' assymetry, it seems much more easier to implement these constraints on the sender's side: to represent a sender who wants to avoid a commitment to having the preferences associated with a given type t_{bad}, it suffices to model the bad consequences of such a commitment, simply assuming an action $a_{t_{\mathsf{bad}}}$ optimal for the receiver iff t_{bad} is the sender's actual state, and which, in that case, comes with a drammatically negative payoff for the sender. Doing the same on the receiver's side, can, at best, only be achieved by refining the set of sender types so as to distinguish between sender types accordingly to (at least some of) the subsequent conversational moves that the sender will perform. But such a trick essentially amounts to consider sequential games with more than a single turn, which we advocate in the next sections.

Let us conclude with a more formal view of the above considerations and an example: how does *deception* in the sense that we encountered in our examples translates in signaling terms? We can define it this way: a sender S of type t, decieves a receiver R if he can induce an action a' from the receiver which is suboptimal for him given $\langle t, m\rangle$, *i.e.* such that there exists another action a verifying $U_R(t, m, a) > U_R(t, m, a')$. Let, for any pair $\langle t, m\rangle$ of type and message, $a^*_{t,m}$ denote the optimal action for a receiver knowing that t is the actual state and m was sent. Following our notion of deception, we say that the type t_{bad} has an *interest to decieve* using message m iff there is a state t_{good} such that we have $U_R(t_{\mathsf{good}}, m, a^*_{t_{\mathsf{bad}},m}) < U_R(t_{\mathsf{good}}, m, a^*_{t_{\mathsf{good}},m})$ and $U_S(t_{\mathsf{bad}}, m, a^*_{t_{\mathsf{bad}},m}) < U_S(t_{\mathsf{bad}}, m, a^*_{t_{\mathsf{good}},m})$.

Let T denote the set of sender types and T_{good} denote the set of "good" sender types, *i.e.* the subset of the set of sender type for which the receiver has better payoff using $a^*_{t_{\text{good}},m}$ than $a^*_{t_{\text{bad}},m}$ upon reception of m: $T_{\text{good}} = \{t \mid U_R(t, m, a^*_{t_{\text{bad}},m}) < U_R(t, m, a^*_{t_{\text{good}},m})\}$.

Fact 41. Using the above notations, we have the following: whenever a rational receiver is decieved by m, *i.e.* he takes $a^*_{t_{\text{good}},m}$ with non zero probability after receiving m, he must believe after reception of m that S being of a "good" type is at least $\frac{\delta_{\text{good}}}{\delta_{\text{bad}}}$ as likely as S being of type t_{bad}, where δ_{bad} is the difference of payoff for the receveir between choosing $a^*_{t_{\text{bad}},m}$ and $a^*_{t_{\text{good}},m}$ in t_{bad}, and $\delta^*_{t_{\text{good}},m}$ is the maximal difference of payoff for the receiver bewteen choosing $a^*_{t_{\text{good}},m}$ over $a^*_{t_{\text{bad}},m}$, in any other state.

As an immediate corollary, it follows that in any perfect bayesian equilibrium with a pure sender strategy sending m in state t_{bad}, if R is decieved then the prior probability of being of a "good" state in which S sends m must be at least $\frac{\delta_{\text{good}}}{\delta_{\text{bad}}}$ as likely as the prior probability of S being of state t_{bad}.

Proof. Let $\rho(\cdot|m)$ be a receiver strategy. Let $\mu(\cdot|m)$ be a probability distribution over sender types such that ρ is rational given belief in μ and such that $\rho(a^*_{t_{\text{good}},m}|m) > 0$. By definition, if ρ is rational given μ, $a^*_{t_{\text{good}},m}$ must be a best response to m. In particular $a^*_{t_{\text{good}},m}$ must yield a better (or equal) expected utility that $a^*_{t_{\text{bad}},m}$ which writes as:

$$\sum_{t\in T} \mu(t|m) U_R(t, m, a^*_{t_{\text{good}},m}) - \sum_t \mu(t|m) U_R(t, m, a^*_{t_{\text{bad}},m}) \geq 0$$

that is to say

$$\left(\begin{array}{l} \displaystyle\sum_{t\in T_{\text{good}}} \mu(t|m) \left((U_R(t, m, a^*_{t_{\text{good}},m}) - U_R(t, m, a^*_{t_{\text{bad}},m}) \right) \\ \displaystyle - \sum_{t\in T\setminus T_{\text{good}}} \mu(t|m) \left(U_R(t, m, a^*_{t_{\text{bad}},m}) - U_R(t, m, a^*_{t_{\text{good}},m}) \right) \end{array} \right) \geq 0$$

Notice that both term of the above difference are positive (the first term on the left is strictly positive since $t_{\text{good}} \in T_{\text{good}}$). Let $\delta_{\text{good}} = max_{t\in T_{\text{good}}}(U_R(t, m, a^*_{t_{\text{good}},m}) - U_R(t, m, a^*_{t_{\text{bad}},m}))$ and $\delta_{\text{bad}} = U_R(t_{\text{bad}}, m, a^*_{t_{\text{bad}},m}) - U_R(t_{\text{bad}}, m, a^*_{t_{\text{good}},m})$. Since $t_{\text{bad}} \in T \setminus T_{\text{good}}$ we have:

$$\sum_{t\in T_{\text{good}}} \mu(t|m)\delta_{\text{good}} - \mu(t_{\text{bad}}|m)\delta_{\text{bad}} \geq \left(\begin{array}{l} \displaystyle\sum_{t\in T_{\text{good}}} \mu(t|m) \left((U_R(t, m, a^*_{t_{\text{good}},m}) - U_R(t, m, a^*_{t_{\text{bad}},m}) \right) \\ \displaystyle - \sum_{t\in T\setminus T_{\text{good}}} \mu(t|m) \left(U_R(t, m, a^*_{t_{\text{bad}},m}) - U_R(t, m, a^*_{t_{\text{good}},m}) \right) \end{array} \right)$$

Hence we must have

$$\sum_{t\in T_{\text{good}}} \mu(t|m)\delta_{\text{good}} - \mu(t_{\text{bad}}|m)\delta_{\text{bad}} = \mu(T_{\text{good}}|m)\delta_{\text{good}} - \mu(t_{\text{bad}}|m)\delta_{\text{bad}} \geq 0$$

and since $\delta_{\text{bad}} > 0$ by hyptothesis we must have $\mu(T_{\text{good}}|m)\frac{\delta_{\text{good}}}{\delta_{\text{bad}}} \geq \mu(t_{\text{bad}}|m)$ which concludes the proof. QED

What this fact shows, is that, in a signalling game, the only basis for a receiver to ever accept a deceptive move (*e.g.* a misleading answer) is that he judges it more likely (modulo the ratio $\delta_{\text{good}}/\delta_{\text{bad}}$ which quantifies the "badness" of the deception) that his opponent is of a "good" type, never that he lacks an argument to confront him, or has reasons to avoid confrontation. Yet we are convinced that the latter are equally good reasons to accept "dodging" moves. Consider as an example the following conversation:

(6.4.1)a. *A*: Are you available on thursday afternoon? I need help moving in.
b. *B*: I have a very important meeting on thursday.

It seems intuitive to assume that *B* might be of two distinct types: the "free on thursday afternoon" type t_f, and the "not free on thursday afternoon" type $t_{\neg f}$. Assume further that *B*'s meeting is scheduled early thursday morning, so that despite telling the truth, *B* is in fact of type t_f (were *B* willing to, he would still be able to help *A* in the afternoon). *A*'s best interest in that case involves intuitively to have *B* commit to being of type t_f as it makes it socially difficult for *B* to refuse his help (and unless *A* has, and is willing to use, another way to pressure *B*, this is likely to be the best he can achieve by talking). Assume finally that *A*, for some reason, has an incentive to not trust *B* and think that it is much more likely that is available than not (because for instance, they work at the same company and *B* has a reputation of being very lazy). Let us have a look at some of the possible responses *A* can make:

c_1 *A* can take *B*'s answer has a no with "*A*: OK, too bad".

c_2 *A* can request a direct answer "*A*: So you are not available on thursday afternoon?"

c_3 Depending on his relation with *B*, *A* can try other, more gentle or less direct denunciations of *B*'s misdirection, for instance we can imagine something in the spirit of "*A*: Oh, I forgot about that! Those morning meetings can be exhausting, I take it that you'll be too tired in the afternoon then?"

Naturally, the best option for *A* depends on his familiarity with *B*, but the point still, is that the payoff of continuation c_2 highly depends on *B*'s reaction. In many cases, *A* might judge c_2 too abrupt, and not really helpful in the case of *B* opting for an explicit *no* answer in return. *A* might thus think that he has little to gain using c_2, and *B*, counting on that, even if he knows that *A* believes him to be more likely available than not, will still try to misdirect, expecting *A* to opt for c_1, or at worse for something like c_3 which, unlike c_2, saves his face.

Indeed, if *A* opts for c_2, *B* might react badly, for instance saying, annoyed, "No, as I just said, I have an important meeting"[27] Arguably this should yield a bad payoff for both *A*, who has not fulfilled his objective[28], and *B* who has been compelled to lie. But importantly, it is also possible that *B* alternatively drops his attempt at misdirecting (*e.g.*, saying "Well, actually my meeting is in the morning, so I guess I'm free after all."), in which case, c_2 should yield high payoff for *A* and low payoff for *B*.

In summary, these considerations show that the rationality of *B*'s attempt at a misleading answer is much more tight to the probability of the different continuations of c_2 to take place than on *A*'s prior belief about *B*'s actual type. Is it possible to account for this using a signaling game? From what we have shown, only at the cost of an "odd", non transparent payoff structure, and/or a rather complicated type-space: let `meeting` denote *B*'s answer to the question. As *B*'s objective is to avoid commiting to t_f, we should have $U_B(t_f, \texttt{meeting}, c_2) < U_B(t_f, \texttt{meeting}, c_1)$. The real struggle is to set the payoff for c_2. If we let $U_A(t_f, \texttt{meeting}, c_2) > U_A(t_f, \texttt{meeting}, c_1)$ then, given *A*'s prior, we are in a case of application of fact 41 and *B* is not rational in trying to misdirect. If we let $U_A(t_f, \texttt{meeting}, c_2) < U_A(t_f, \texttt{meeting}, c_1)$, then an important part of the reasonning becomes hidden in the payoffs, and the games becomes "artificially" cooperative: any explicit encoding of the opposing interrests at stake is left behind.

Again, a solution to this dilemma, is to refine the set of types distinguishing for instance between *B*'s of type t_f that reacts agressively to c_2 with the full lie and those that rather drop their misdirection attempt.

[27] Notice also that, interestingly, even if *B* with such an answer, explicitely lies about his availability, he remains only implicitely commited that it is the meeting that makes him unavailable. So he can still drop this commitment at the cost of admitting that he was incoherent or not responsive. Hence, even if *A*, for some reason, is willing to confront *B* for lying to him, and has formal evidence that the meeting is indeed in the morning, doing so still requires a lot of efforts on his part.

[28] unless of course *A* is ready to accuse *B* of lying to him.

But this mean that we must condition sender strategies to their type, and more generally this boils down to considering another class of games that are involve more than one sending and response.

Other models like that of Glazer and Rubinstein (2004) exist that do not use signaling games. However, they also have difficulties in expressing the sort of constraints we have developed above. Signaling games and persuasion games both still take a broadly Gricean view of communication: conversations are essentially information gathering or exchange activities; agents exchange messages for the purpose of affecting the beliefs of the other partner. This is precisely, however, what is in doubt in many conversations. In many conversational settings and in all of our examples, agents converse not in the hope persuading their opponents, but rather to impress or persuade others, and perhaps themselves. Just as Grice captures important aspects of some but not all conversations, people do try sometimes to persuade or to exchange information, but this is not a general framework for all conversations.

Chapter 7

Message-exchange games

Contents

7.1 Switching to sequential games

Following the conclusions of the previous chapter, we need a different (more general) model of conversation departing from signaling- or other one-shot games models of the litterature. We still want to represent conversations as games in which the players are trying to achieve a certain end–namely, that the conversation go in a particular way, but we want to adopt an **explicit** representation of agents divergent goals without commiting to a complete lack of communication. To this aim, we now delve deeper into the structure of conversational games. What do they concretely involve? What are the 'moves' of the players, what are their

'strategies' and so on? What are their winning conditions? And how can we model conversational goals in a formal setting?

Recall feature (I-IV) from section 6.3.2. In order to turn these into a model, we need three things:

(i) an appropriate vocabulary of conversational moves for building sequences of message exchanges between players,

(ii) goals or winning conditions for conversational players and

(iii) a way of modeling the epistemic limitations that players cannot in general foresee the last move of a conversation.

Infinitary games like Banach Mazur (BM) games (Oxtoby, 1957; Grädel, 2008; Kechris, 1995) furnish a good point of departure, as they reflect some features of (I-IV). For simplicity, we will mostly restrict our attention to two-player win-lose games, allowing us to concentrate on basic conceptual points. These games involve a set of sequences of conversational moves and a characterization of winning conditions for players of the game. We modify them however, in order to distinguish between turns of talk so as to keep track of "who said what", an information which is crucial in many settings. Our theory thus distinguishes between conversations in virtue of their winning conditions, and different winning conditions require different strategies for achieving them giving rise to different linguistic realizations. We show the game's determinacy and discuss existence of these strategies in various cases. We introduce a formal model of winning conditions, enabling us to compare different conversational goals and their winning strategies. Finally, we propose a formalization of *misdirection* and *blindness* to explain the rational basis of engaging in o-sum conversations.

7.2 What do agents communicate in strategic context?

In the pictures we have sketched so far, players interact with each other and exchange messages in a sequential process. But what do these messages convey, and how do they help players achieve their objectives?

Our foundational hypothethis, is that messages convey objective, public commitments. For instance, D in (6.3.4-d) commits to Dr. Tzeng's having "torn apart" the nerve by agreeing with LP's description in (6.3.4-c). D then tries to go back on that commitment in (6.3.4-g). D may or may not believe this commitment. But if he signal agreement with (6.3.4-c), then he is committed to its content, and he can be attacked on the basis of that commitment or subsequent commitments. As the set of examples of section 6.3.1 illustrates, conversationalists often pay careful attention to the commitments of others, which encompasses not only explicit commitments but also to their implicatures. In example (6.3.5) for instance, Bentsen seizes on a weak or possible implicature of Quayle's commitments, that he is comparable in Presidential stature to JFK, and attacks Quayle on that basis. The moves players make to defend their commitments or to attack those of an opponent exploit the conventional meanings and even the implicatures that messages have. So an appropriate model of the conversational arena must enable us to fix the meanings of players' moves to their conventional meaning.

Why do conversationalists make the commitments they do, if they don't do it to persuade their interlocutors or to send a signal that their interlocutors will find credible? Players make the commitments they do, for the purpose of convincing or influencing an idealized third party, which we call *the Jury*. The Jury is for us an abstract role that can be satisfied in diverse ways. In some cases the role of the Jury might be endorsed by an actual agent, taking part or simply listening to the conversation, *e.g.* in examples (6.3.2) and (6.3.4), the jury of the court or in examples (6.3.5) and (6.3.7), the American electorate. Sometimes it can even be endorsed by one of the players, as in example (6.3.3). As such, the Jury does not perform

conversational moves but is rather a scoring function for the game (even when it falls to one of the conversational agent to play the role of the Jury, we conceive of these two roles separately–as a player she can make move, as the Jury she can only evaluate). Players choose their conversational objectives based on what they believe they can defend against their opponents and that will find favor in some way with the Jury. A player attempts to achieve her conversational objective, while her opponent tries to thwart her. The Jury is an abstraction of a rational and competent user of the language of the players and judges on that basis whether a given discourse move or a sequence of moves contributes toward the realization of the conversational objectives of a player or not.

Let us now conclude this preamble to the fleshing of the model in full detail with a discussion of a last, yet essential, characteristic of the model: the view of conversations as **infinite** sequences of moves.

7.2.1 Why infinite games?

As principle (IV) of section 6.3.2 already hints at, we believe, that humans must act as though conversational games were unbounded. If conversations have definite last moves and our players have opposing interests, even the presence of the Jury will not explain why our agents converse in the way they do.

Consider example (6.3.2) again, or its non courtroom variant, example (6.3.3). Janet is presented with a Hobbsian choice. Ideally, she would prefer not to answer the question at all or simply lie. To not answer the question or to lie would be rational and what Janet should do, if she were playing a one shot game with no further interaction with Justin (this is akin to the defect move in the Prisoner's Dilemma). As conversations, however, have continuations, many people have the intuition that a refusal to answer will make Janet fare worse in subsequent exchanges. Janet cooperates with her interlocutor in the minimal sense of providing a response to the question, what Asher and Lascarides (2012) call *rhetorical cooperativity*, because of reputation effects. If Janet doesn't cooperate by responding to Justin, she risks receiving uncooperative treatment if in the future she asks a question or make some demands of him. This is a form of the "tit for tat" view of Axelrod (2006). Nevertheless, the reputation argument has its problems. If a conversation is just a finite sequence of one shot games, what holds for a one shot game holds throughout a conversation. Backward induction over such a finite sequence would lead Justin to the conclusion that he should not bother to ask his first question because it is in Janet's interest to defect at the earliest possible opportunity. If there is a foreseeable last move for one of the players i, then she will play to her advantage and defect on the last move, if her opponent has gone along in the discussion. The opponent seeing this will reason by backwards induction to defect at the earliest possible moment. The prediction is that given a foreseeable last move, no message exchange should occur.

If the conversational game is assumed to be infinite, however, the formal argument for the rationality of defection over sequences of exchanges in cases where conversationalists have opposing interests disappears. The argument from backward induction fails because there is no last move from which to begin the induction. However, there is still some explaining to do. A simple "tit for tat" model doesn't explain why interlocutors cooperate with each other rhetorically, *even if* their roles *vis a vis* their interlocutors are never reversed, even if Janet and Bronston never make any demands of their interlocutors.

In our model, the Jury can force rhetorical cooperativity. A defection will hurt player i if the Jury can infer that i is defecting because a rhetorically cooperative move would reveal a reason for them not to be persuaded by her. This is also a feature of the model of Glazer and Rubinstein (2001, 2004) but their model is more restrictive. In their model, the Jury only interacts with one sender who must persuade the Jury to accept or reject a message. In addition, the sender of a message is restricted in her choice of messages she can send in a given state, and so the Jury can draw more secure inferences from messages she doesn't send. Since she can only send certain messages in certain states (e.g., Bronston might be able to say he did not have an account only in a state where he truly does not), a failure to send a message or to respond to a question where the message is directly requested and would be in the player's interest to send could well

indicate that the player is not in the state where such a message is permitted.

We have made no such assumptions about messages, however, because we do not think that messages are tied to states in such a simple way. One can say anything regardless of the state of the world in a conversation (point (III) of our conversational features in section 6.3.2). So the reasoning from signals and strategies to the persuasiveness of a player is much more uncertain for the Jury in our games. In addition, while Player i needs to convince the Jury that she has achieved her conversational goals, her goals are more complex than simply getting the Jury to accept the content of a particular message and crucially involve her opponent. Player i could simply refuse to cooperate with her opponent, because she has a general strategy of not revealing information to her opponents. Or she could provide a reasonable defense for why she is not cooperating. In either case, it falls on *the opponent* to make the case to the Jury that player i's lack of rhetorical cooperativity provides a reason to reject i's goals. If attacked, player i can reply to the opponent, defending her lack of cooperativity, and then the opponent must press the issue.

The following excerpt from a press conference by Senator Coleman's spokesman Sheehan brings out these features of our model. Senator Coleman was running for reelection as a senator from Minnesota in the 2008 US election (we thank again Chris Potts for bringing this example to our attention):

(7.2.1)a. Reporter: On a different subject is there a reason that the Senator won't say whether or not someone else bought some suits for him?

b. Sheehan: Rachel, the Senator has reported every gift he has ever received.

c. Reporter: That wasn't my question, Cullen.

d. Sheehan: The Senator has reported every gift he has ever received. We are not going to respond to unnamed sources on a blog.

e. Reporter: So Senator Coleman's friend has not bought these suits for him? Is that correct?

f. Sheehan: The Senator has reported every gift he has ever received. (Sheehan continues to repeat "The Senator has reported every gift he has ever received" seven more times in two minutes to every follow up question by the reporter corps).

The record is available online at: `http://www.youtube.com/watch?v=VySnpLoaUrI`.

Sheehan, like Bronston in example (6.3.2), is seeking to avoid committing to an answer to a question. Sheehan's move (7.2.1-b) in response to the reporter's first question could be interpreted as an indirect answer, a statement that implicates a direct answer; the senator did not comment on the question concerning whether he had received the gift of suits because he felt he had already said everything he had to say about the matter. But in example (7.2.1) the reporter does not accept this indirect answer; she says that Sheehan's response is not an answer to her question. In effect, she wants a direct answer to the question concerning the suits. Sheehan with move (7.2.1-d) then explains why he won't answer the question. The reporter then presses the issue, and Sheehan becomes rhetorically uncooperative for the rest of the exchange, repeating the same thing. At this point, the Jury will begin to reflect on Sheehan's strategy: is he being rhetorically uncooperative because he has something to hide? His earlier explanation for his defection from rhetorical cooperativity becomes lost, and it becomes more and more plausible that Sheehan won't answer the question because the true answer is damning to his interests. To win given a defection from rhetorical cooperativity, Sheehan has to have a reply for every attack; on the other hand, if the opponent eventually introduces an attack for which the first player does not have a convincing reply (e.g., he simply repeats himself or simply stops talking), the opponent will win.

The need to justify uncooperative moves or defection generalizes. In most strategic situations, in order to win, o must engage with questions and remarks of her opponent(s); she must show that her opponent cannot attack her position in such a way that a rational unbiased bystander would find plausible. For any discourse move, we can imagine a potential infinity of attacks, defenses and counterattacks. In successful play, a player has to be able to defend a move m against attacks; she may have to defend her defense of

m against attacks and so forth. This is a general necessary victory condition for o. Let $attack(n, m)$ hold iff move m attacks move n; commitments or types of discourse moves that generalize over more specific discourse moves[29] that are used to defend or attack commitments:

Observation 1. [NEC] A play is winning for o only if for all moves n of o and for all moves m of 1, $\text{attack}(n, m) \rightarrow \exists k(\text{move}(o, k) \wedge \text{attack}(m, k))$

Conversely for 1, a sufficient condition for winning is the negation of Observation 1. Given (1) o wins only if she is prepared for the conversational game never to end and to rebut every attack by 1. It is this constraint that provides a second reason for assuming conversational games to be infinite and is a powerful reason for obeying rhetorical cooperativity.

NEC also has empirical consequences. In virtue of it, we can see why Quayle intuitively loses example (6.3.5). Part of Quayle's winning condition was not to come under attack for his implicit comparison or at least to be able to rebut any attack on his move; that is NEC was also part of his winning condition. But given that he had no rejoinder to Bentsen's unanticipated move, he failed to comply with Observation 1 and so lost.

To Observation 1, we add another, motivated by example example (7.2.1): to win, o should not simply repeat herself in the light of a distinct move by her opponent at least not more than twice.

Observation 2. [NR] A play is winning for o only if there is no move k by o such that o repeats k on m successive turns, for $m \geq 3$, regardless of what 1's intervening contributions are. This observation buttresses NEC's support for rhetorical cooperativity.

While we could weaken the quantifiers in NEC and NR to something like *for most moves of o*, we are rather interested in the general upshot of such constraints: to model winning play by o, we need to model a conversation as a potentially unbounded sequence of discourse moves, in which she replies to every possible attack by her opponent. Moreover, at least some of the moves of player i must be related to prior moves of her opponent. It follows that it is always risky, and often just rationally unsound, to play a rhetorically uncooperative move like defection without further explanation that is optimal only if it is the last move in a finite game. Defection from rhetorical cooperativity is possible, but it must be explained or defended in any winning play convincing the Jury. A player who plays a rhetorically uncooperative move opens himself up to an attack that will lead to a defeat in the eyes of the Jury, as in example (7.2.1). That is, relatively weak and uncontroversial assumptions about the beliefs and preferences of the Jury validate rhetorical cooperativity as a component of any winning play.

Even if in practice, conversations don't go on forever, players have to worry about continuations of conversations thus should rationally act as if a conversation were 'potentially infinite'. In such situations, a theory of finite play does not apply and one has to resort to infinite plays. This is why it is necessary to adopt a framework of infinite games. By moving to a framework with unbounded conversational sequences, Aumann and Hart (2003); Aumann and Maschler (1995) show how games with unbounded cheap talk, games involving extended conversations with an infinite talk phase consisting of a pattern of revelations and agreements ending ultimately in an action, make possible equilibria for players that are not available in one shot or even sequences of revelations of bounded length. While we have adopted the simplest of payoff structures for our study, our examples show that unbounded conversational sequences allow players to win conversations that they otherwise couldn't. Had the reporters in example (7.2.1) been limited to one question and one follow up, they could not have successfully attacked Sheehan in the way they did.

Another reason in favor of using infinitary games is, paradoxically, their simplicity. Given that we cannot impose any intrinsic limitations as to the length of conversations, a formalization of purely finite

[29] Examples of such moves are Answering a question, Explaining why a previous commitment is true, Elaborating on a previous commitment, Correcting a previous commitment, and so on–in fact, they correspond to the discourse relations of a discourse theory (Asher and Lascarides, 2003).

conversations is more complicated. In an infinitary framework, it is also straightforward to model finite conversations. Finite conversations are not just conversations that stop but crucially involve a point of mutual agreement that the players have finished (Sacks, 1992). We represent a finite conversation then as one in which a finite sequence terminates with an agreement on a special "stop" symbol that is then repeated forever. More than that, initial prefixes of infinite sequences will play a very important role in the sequel. While our models of conversations will be infinite sequences, all that we ever make judgments on are finite prefixes of such conversations. We will have to evaluate the play of players and whether they have yet met or are meeting their objectives on such finite prefixes. Importantly though, evaluating a finite prefix and declaring a definitive winner are two different things, and meeting one's objective on a finite prefix is not always sufficient. Typically one has to ensure a form of *stability* in the way one achieves one goal.

7.3 Message Exchange Games defined

We have established that conversations should be modeled as some sort of infinite games. In this section we define such games, which we call *message exchange games* (ME games). The first step we take, is to define what a *message* his, *i.e.*, to set a generic vocabulary for ME games.

7.3.1 The vocabulary of discourse moves

Chapters 2 and 5 have insisted upon the fact that messages in a conversation (or a discourse) are typically related to each other. For instance, in example (6.3.2), (6.3.2-b) is related to (6.3.2-a) via an answerhood relation, and (6.3.2-c) is related to the pair of (6.3.2-a) and (6.3.2-b) as some sort of follow up question. Furthermore, this links come with semantic consequences which are essential to determine whether a player has come closer to achieving her goals or not (and vice versa), and how he must respond and link to his interlocutors contribution to win. The fact that (6.3.2-d) at least implicates an answer to (6.3.2-c) means that if the Prosecutor's goal is merely to get an implicated commitment, then he has at that point achieved his victory condition.

Linguistic theories of discourse structure like SDRT, introduced in section 2.3, have developed a rich language to characterize the semantics and pragmatics of moves in dialogue. We have seen that SDRT features discourse labels, representing *variables* for dialogue moves. These moves are characterized by contents that they commit their speaker to, and crucially some of this content involves predicates that denote coherence relations between moves–like the relation of *question answer pair* (`QAP`), in which one move answers a prior move characterized by a question.

We will therefore rely on SDRT's concepts to set up our vocabulary: the syntax for dialog moves that we adopt is a simplified and linearized version of the syntax of SDRSs. This choice allows us to talk about discourse moves and coherence relations while keeping the process of compositionally merging a sequence of moves into a formal representation of the conversational context straightforward, which our games require. In particular, we suitably abstracts over some peculiarities of SDRT's complex syntax-semantics interface. As we focus in this chapter on the exposition of the game model itself, and since the semantics of discourse moves constitutes a rather vast topic on its own, we dedicate another part of the present thesis, part III, to a detailed discussion of the logical language, its semantics and important phenomena such as ambiguity, acknowledgments and a semantic definition of *attacks*.

For our present need, it will be sufficient to assume that the elements of syntax of our vocabulary contains a countable set of discourse constituent labels $DU = \Pi_0 \cup \Pi_1 = \{\pi, \pi_1, \pi_2, \ldots\}$, partitioned between labels of player 0 (Π_0) and labels of player 1 (Π_1), a finite set of discourse relation symbols $\mathcal{R} = \{R, R_1, \ldots R_n\}$, and formulas $\varphi, \varphi_1, \ldots$ from some fixed language for describing elementary discourse move contents. Our generic vocabular $V_{\mathcal{U}}$ consists in a set of *discourse moves* or *messages*, which are formulas of

the form $\pi : \chi$, where χ is a description of the content of the discourse unit labeled by π, *i.e.* either a formula φ of the fixed language for the content of elementary unit, or a formula of the form $R(\pi_1, \pi_2)$, which says that π_1 stands in coherence relation R to π_2. This settings naturally inherits SDRT's notion of complex constituent (see section 2.3): we say that a formula $\pi : R(\pi_1, \pi_2)$ makes π_1 and π_2 subconstituents of π. We say that π *labels* the move $\pi : \chi$. We let $V^0_{\mathcal{U}}$ and $V^1_{\mathcal{U}}$ denote the sets of moves labeled by labels in Π_0 and Π_1 respectively. Following Asher and Lascarides (2003), each discourse relation symbolized in $V_{\mathcal{U}}$ comes with constraints as to when it can be coherently and consistently used in context and when it cannot.

We have assumed Π_0 and Π_1 to be disjoint, so that each move's label encodes the information of whether agent 0 or 1 performed the move. However, moves of $V^{\mathcal{U}}$ are SDRSs (or simplified SDRSs), and as such intermediary mental representations (see section 2.3) that we assume are derivable from more elementary representations (*e.g.*, the sequence of syntactic representations of EDUs' content). So, at a lower level of analysis, 0 and 1, do not choose their contributions in distinct vocabularies, they both say things, for instance in English or French, and in principle, they can say anything. Only in the process of deriving an SDRS for a given move in its context, the choice of labeling with a label of 0 or a label of 1 will be made, accordingly to who performed the move. In ME games, assuming a level of analysis where players moves are already SDRSs will help keep things as simple as possible. Yet it will proove useful on some occasion to keep in mind that, at a lower level of analysis, the distinct vocabularies might be identified in such a way that given a particular play, including information about turn-taking, we recover the separate vocabularies representations. The simplest way to achieve this while still abstracting over the complexities of the syntax/semantics interface, is to simply assume that $\Pi_o = \Pi \times \{i\}$ for a common set of anonym label symbols Π. The vocabulary $\tilde{V}^{\mathcal{U}}$ of "anonym" moves is defined as the set of moves of the form $\pi : \chi$ where χ is exactly as before (thus, possibly involving relational predicates refering to "attributed" labels in Π_0 and Π_1) but $\pi \in \Pi$ is an anonym label. It is straightforward to see that moves anonymization yield two bijections, between $V^{\mathcal{U}}_0$ and $\tilde{V}^{\mathcal{U}}$, and between $V^{\mathcal{U}}_1$ and $\tilde{V}^{\mathcal{U}}$ respectively. Thus, the function anonymizing all moves in a sequence of moves by 0 and 1 is a non-injective surjection from sequences of moves in $\tilde{V}^{\mathcal{U}}$ to sequences of anonym moves. Only **given a particular decomposition of a sequence of anonym moves into a structure of turns**, we might reattribute moves' labels.

ME game messages come with a conventionally associated meaning in virtue of the constraints enforced by the Jury; an agent who asserts a content of a message *commits* to that content, and it is in virtue of such commitments that other agents respond in kind. While SDRT has a rich language for describing dialogue moves, it is not explicit about how dialogue moves affect the commitments of the agents who make the moves or those who observe the moves. part III summarizes the results of Venant et al. (2014); Asher and Venant (2016); Venant (2015) linking the semantics of the SDRT language with commitments explicitly (in different ways). We will augment the SDRT language with formulas that describe the commitments of dialogue participants: where φ is a formula for describing a content, $C_i\varphi$ is a formula that says that player i commits to φ. Commitment is modeled as a Kripke modal operator via an alternativeness relation in a pointed model with a distinguished (actual) world w_0. Which grants a semantics for discourse moves that links the making of a discourse move by an agent to her commitments: i's assertion of a discourse move φ, for instance, we will assume, entails a common commitment that i commits to φ, written $C^*C_i\varphi$, where $C^*C_i\varphi$ entails arbitrary nestings of $C_{x_0} \ldots C_{x_k}\varphi, k \in \omega, x_i \in \{0, 1\}$. We will also show each discourse move $\pi_i : \varphi$ defines an action, a change or update on the model's commitment structure; for a propositional language, this can be done in the style of public announcement logic viz. Baltag et al. (1999); Baltag and Moss (2004). For instance, if agent i asserts p, then the commitment structure for the conversational participants is updated such so as to reflect the fact that C^*C_ip. Postponing detailed definitions and a thorough study of the semantics of commitments, let us simply assume in the following that we dispose of a class of models and interpretation functions $\|\mathfrak{A}\|^{\cdot}$ that allows us to interpret a discourse move x as an update of the model $\mathfrak{A}$ into a new model $\|\mathfrak{A}\|^x$. This semantics is useful because it allows us to move from sequences of discourse moves to sequences of updates on any model for the discourse language. So

a sequence of moves $x_0.y_1$ in a model with commitments for 0 and 1 will yield given an input model $\mathfrak{A}$, a sequence of models $\mathfrak{A}^{\|x_0\|}$, $(\mathfrak{A}^{\|x_0\|})^{\|y_1\|}$.

To conclude on the vocabulary and its interpretation, consider example (6.3.2) again. Associated syntactic representations are given below:

Speaker	Sentence	Formula in $\mathcal{L}$
P	a.	$\pi_A : ?\,\mathsf{bank}$
B	b.	$\pi_B : \neg\mathsf{bank} \wedge \pi_B : \mathsf{QAP}(\pi_A, \pi_B)$
P	c.	$\pi_C :?\mathsf{bank\_ever} \wedge \pi'_C : \text{Q-ELAB}(\pi_A, \pi_C)$
B	d.	$\pi_D : \mathsf{company} \wedge \pi'_{ans} : \mathsf{IQAP}(\pi'_C, \pi_D)$

Where $\pi_A, \pi_C, \pi'_C \in \Pi_0$, while $\pi_B, \pi_{ans}, \pi'_{ans} \in \Pi_1$.

7.3.2 Definition of ME games

Now that we dispose of a (generic) vocabulary of discourse moves, we are ready to define message-exchange games. We use Banach Mazur games, a well-known sort of infinitary game, as a departure point and point of comparison. We then make some remarks about the expressive capacities of our new framework and examine how it addresses the problems we found with the signaling game framework.

A Banach Mazur game involves a countable, non-empty set vocabulary vocabulary V. For linguistic modeling, we will generally assume that $V \subseteq \tilde{V}^{\mathcal{U}}$ For any set X we let X^* denote the set of finite strings over X and X^ω the set of countably infinite strings over X. Let $\cdot$ denote strings concatenation, and for two sets of strings W, W' over some set X let $W \cdot W' = \{w \cdot w' \mid w \in W, w' \in W'\}$. For any $w \in X^*$, $w \cdot W'$ is syntactic sugar for $\{w\} \cdot W'$.

Definition 44. A Banach-Mazur game (BM game) $BM(V^\omega, Win)$ consists of an infinite set of strings V^ω together with a winning condition $Win \subseteq V^\omega$.

The game proceeds as follows. Player 0 first chooses a *non-empty finite* string $x_0 \in V^*$. Player 1 responds by choosing another non-empty finite string $x_1 \in V^*$. Player 0 moves next choosing another finite string x_2. This process repeats itself forever yielding a play, an infinite sequence of alternating moves by 0 and 1. Define the flattening *flat* of a play $p = (x_k)_{k\in\mathbb{N}}$ as the infinite sequence eventually designed by the two players: $flat(p) = x_0 \cdot x_1 \cdot x_2 \ldots \in V^\omega$ where $\cdot$ denotes string concatenation. Player 0 wins the game if $flat(p) \in Win$. Player 1 wins otherwise. A *strategy* f_i for player i, is a function from the set of finite plays to the set of finite strings, V^*. A play $p = (x_k)_{k\in\mathbb{N}}$ of the game, is said to be *consistent* with the strategy f_i iff, for every integer k, $k \bmod 2 = i \Rightarrow x_k = f_i(x_0 \cdot x_1 \cdot \ldots \cdot x_{k-1})$. In other words, each move of player i is played according to f_i. A strategy f_i is said to be *winning* iff in every play consistent with f_i, player i wins.

BM games suggest a natural model for conversations: participants alternate turns in which they utter finite contributions. These contributions add to each other, and together form a conversation. This process potentially goes on indefinitely, or, at least strategic reasoning requires thinking of it that way. However, BM games "erase" the information of who said what in the following sense:

Proposition 1. Let $BM(V^\omega, Win)$ be a BM-Game. Then, for any play ρ and every play $\rho' \in flat^{-1}(flat(\rho))$, player i wins ρ iff player i wins ρ'.

Given any infinite sequence s, any infinitely countable set $turns \subseteq \mathbb{N}$ such that $min(turns) > 0$ yields a play in $flat^{-1}(s)$ and conversely: every element in $turns$ specifies a position in s which is the end of a player's move. A corollary of the above proposition is that we cannot define a winning condition that imposes for instance that player 1 says something in particular, as long as she and 0 don't infinitely repeat the same single move. We formalize this observation as follows:

Corollary 9. Let $BM(V^\omega, Win)$ be a BM-Game. Let $y \in V^*$ be a finite sequence such that there is at least one infinite sequence w in Win such that $w \notin V^* y^\omega$ or $|y| > 2$. There is a play ρ of the game such that y is never a substring of any move of player 1 in ρ.

The idea for the proof of this corollary is simple: since w does not end with infinite repetitions of y, every occurrence of y in w is eventually followed by something which is not y, call it x. It suffices to define the alternation of turns so that x constitute exactly the turns of 1. More formally: let i be a position in w at which y appears. define l_i as $Min(\{l \in \mathbb{N} \mid y \text{ does not occur at } i + l \times |y| \text{ in } w\}$. l_i exists by hypothesis since otherwise w would end with infinite repetitions of y. For any position k in s, let $n_y(k)$ be the position of the first occurence of y in w after k. Define inductively *turns* with $n_y(0) + l_{n_y(0)} \times |y| \in turns$, and for any k in turns, $n_y(k) + l_{n_y(k)} \times |y| \in turns$. *turns* yields a winning play in $flat^{-1}(w)$ for which player 1 never "says" y.

BM games have a limitation that require us to introduce a more structured type of game given our principle (I). A given conversationalist might have as a goal that her interlocutor *and only her* commits to a particular content, or answer a particular question, which BM games do not allow. Consider again example (6.3.2). There's an important difference between Bronston's response to a question by the prosecutor and the prosecutor's offering that information himself, and that difference can't be captured in BM games under all interesting scenarios. Discourse moves contain more information than the sentence itself. The discourse move that Bronston commits to a negative answer to (6.3.2-a) provides more information than just the string *no sir* provides, and it is such moves that are of interest.

To remedy the expressive limitations of BM games, we introduce our variant, message exchange (ME) games:

An ME game involves two disjoint vocabularies of discourse moves, one vocabulary V_i for each player i, $i \in \{0, 1\}$. In some occasion, it will proove usefull to consider the special case of the two players playing with a common set of moves V, *i.e.* the case $V_i = V \times \{i\}$. Assuming disjoint vocabularies allows keeping track of the turn structure. Applying ME games to model linguistic exchanges, we will generally assume, as we did with BM games, that the set of moves V_i is a subset of the generic set of moves of section 7.3.1, and that in addition V_i only consists of discourse moves labeled with a discourse label $\pi \in \Pi_i$.

An ME game is an infinite game where the players 0 and 1 alternate by playing finite sequences of moves from V_0 and V_1. As opposed to BM Games, we do not impose ME games in their most general form to be 0-sum:

Definition 45. An ME game is a tuple.

$$\mathcal{G} = \mathrm{ME}((V_0^+ \cdot V_1^+)^\omega, Win_0, Win_1)$$

with $Win_0, Win_1 \subseteq (V_0^+ \cdot V_1^+)^\omega$.

The game is structured so as to encode the information on which player played what into the plays, and proceeds as follows: a *turn* by i is a non empty finite sequence of elements in V_i. The game (that we will also refer to as *a converstation*) starts with 0 playing an initial turn. Then 1 plays a turn, then 0 plays a turn, and so on and so forth in a strictly alternating fashion. The games ends after an infinite number of steps. Thus, a partial play is a finite sequence of turns in $(V_0^+ \cdot V_1^+)^*$. Define a *round* as a pair containing a turn by 0 and a subsequent turn by 1. A complete play of the game ρ is an infinite sequence of turns which, clustering turns into rounds, can always be uniquely represented as an infinite sequence in $(V_0^+ \cdot V_1^+)^\omega$. Player 0 wins if $\rho \in Win_0$ and Player 1 wins if $\rho \in Win_1$.

Note that it might be the case that $Win_0 \cap Win_1 \neq \varnothing$. Then if $\rho \in (Win_0 \cap Win_1)$ then both players win. Furthermore, Win_0 and Win_1 need not exhaust $(V_0^+ \cdot V_1^+)^\omega$. In that case if $\rho \notin (Win_0 \cup Win_1)$ then neither player wins. Thus we define the following subclasses of ME games.

Definition 46. Let $\mathrm{ME}((V_0^+ \cdot V_1^+)^\omega, Win_0, Win_1)$ be an ME game. Then

- if $Win_0 = \overline{Win_1}$ we say that $\mathrm{ME}((V_0^+ \cdot V_1^+)^\omega, Win_0, Win_1)$ is zero-sum. In that case we denote $\mathrm{ME}((V_0^+ \cdot V_1^+)^\omega, Win_0, Win_1)$ simply as $\mathrm{ME}((V_0^+ \cdot V_1^+)^\omega, Win_0)$,
- if $(Win_0 \cap Win_1) \neq 0$, we say that $\mathrm{ME}((V_0^+ \cdot V_1^+)^\omega, Win_0, Win_1)$ is non zero-sum.

A strategy σ_0 for player 0 in the game, is a function from finite sequences of turns of even length to V_0^+. A strategy σ_1 for player 1 is a function from finite sequences of odd length to V_1^+. We say that a play $\rho \in (V_0^+ \cdot V_1^+)^\omega$ conforms to a strategy σ_i of player i if i plays according to σ_i to generate ρ. A strategy σ_i is winning for player i if every play ρ that conforms to σ_i is in Win_i. Thus, no matter what the other player plays, as long as i sticks to σ_i she wins. A zero-sum ME game $\mathcal{G} = \mathrm{ME}((V_0^+ \cdot V_1^+)^\omega, Win)$ is said to be determined if either Player 0 or Player 1 always has a winning strategy.

For the case where there is a common underlying vocabulary V, *i.e.* where we have $V_i = V \times \{i\}$, we introduce the following definitions which will help us switch between different sets of sequences. We let π denote the natural projection of $V_0 \cup V_1$ onto V ($\pi(v, i) = v$). π naturally extends to a function $\pi_\star$ over finite sequences in $(V_0 \cup V_1)^*$: $\pi_\star(x_0 \dots x_k) = \pi(x_0) \dots \pi(x_k)$, and to a function π_ω over infinite sequences in $(V_0 \cup V_1)^\omega$: $\pi_\omega((x_k)_{k\in\mathbb{N}}) = (\pi(x_k)_{k\in\mathbb{N}})$. Finally define $\pi_{f,\omega} = \pi_\omega \circ \mathit{flat}$ as the function flattening and projecting sequences of $(V_0^+ \cdot V_1^+)^\omega$ onto V^ω: $\pi_{f,\omega}((\bar{\nu}^k)_{k\in\mathbb{N}}) = \pi(\bar{\nu}^0)_0 \dots \pi(\bar{\nu}^0_{|\bar{\nu}^0|}) \dots \pi(\bar{\nu}^k_0) \dots \pi(\bar{\nu}^k_{|\bar{\nu}^k|}) \dots$ (where all the $\bar{\nu}^k$ are rounds in $V_0^+ \cdot V_1^+$).

To go further into the theory of ME games (or BM games), we need to introduce some topological notions.

Let X be a non-empty set. The Cantor topology on X^ω is defined as follows. A basic open set is of the form xX^ω and is denoted $\mathcal{O}(x)$. The open sets are thus of the form AX^ω where $A \subset X^*$ is a set of finite sequences over X. The closed sets are as usual, complements of the open sets. The sigma algebra generated from the open sets is called the Borel algebra over X^ω, $\mathcal{B}(X^\omega)$. The sets in $\mathcal{B}(X^\omega)$ can also be defined hierarchically as follows.

Let Σ_1^0 be the set of all open sets. $\Pi_1 = \overline{\Sigma_1^0}$ is the set of all closed sets. Then for any $\alpha > 1$ where α is a successor ordinal, define Σ_α^0 to be the countable union of all $\Pi_{\alpha-1}^0$ sets and define Π_α^0 to be the complement of Σ_α^0. For a limit ordinal η, $1 < \eta < \omega_1$, Σ_η^0 is defined as $\Sigma_\eta^0 = \bigcup_{\alpha<\eta} \Sigma_\alpha^0$ and $\Pi_\eta^0 = \overline{\Sigma_\eta^0}$. For every ordinal α, $\Delta_\alpha^0 = \Sigma_\alpha^0 \cap \Pi_\alpha^0$. The infinite hierarchy thus generated is called the Borel hierarchy. Figure 7.1 presents a schematic picture of the initial sets and their inclusion relations in the Borel Hierarchy.

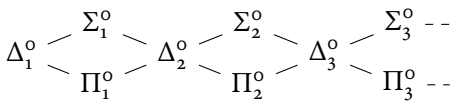


Figure 7.1: The Borel hierarchy

To precisely characterise the complexity of a set $A \subset X^\omega$ we need the notion of completeness which is defined in terms of Wadge reductions. Given vocabularies X and Y a set $A \subset X^\omega$ is Wadge reducible to a set $B \subset Y^\omega$, denoted as $A \leq_W B$, if and only if there exists a continuous function $f : X^\omega \to Y^\omega$ such that $A = f^{-1}(B)$. If $A \leq_W B$ and $B \leq_W A$ then A and B are Wadge equivalent, denoted $A \equiv_W B$. It is known that if $A \equiv_W B$ then A and B belong to the same level of the Borel hierarchy. A set $A \subset X^\omega$ is said to be Σ_α^0-hard (resp. Π_α^0-hard if $B \leq_W A$ for all $B \in \Sigma_\alpha^0$ (resp. $B \in \Pi_\alpha^0$). A is said to be Σ_α^0-complete (resp. Π_α^0-complete) if $A \in \Sigma_\alpha^0$ (resp. $A \in \Pi_\alpha^0$) and A is Σ_α^0-hard (resp. Π_α^0-hard). Intuitively, complete sets for a class represent the most complex sets for that class in terms of structure. If A is Σ_α^0-complete or Π_α^0-complete, we shall say that its Borel complexity is α.

A subset Y of X^ω is nowhere dense if the closure of Y contains no non-empty open set. Y is meager or *topologically small* if it is a countable union of nowhere dense sets. Y is co-meager or *topologically large* if its complement is meager.

A Gale-Stewart game (GS game) $\mathrm{GS}(X^\omega, \mathit{Win})$ (Gale and Stewart, 1953), over a non-empty set X is similar to an ME game where the players take turns in playing 'single' elements from X. *Win* is a subset of X^ω and similar to an ME game Player 0 wins a play if and only if $\rho \in \mathit{Win}$. Martin (Martin, 1975) showed that all GS games where *Win* is a Borel subset of X^ω are determined.

To later define and discuss about "misdirection" we shall need the result that our ME games are determined as well. Hence, relying on Martin's result, we show that

Theorem 1. A zero-sum ME game $\mathcal{G} = \mathrm{ME}((V_0^+ \cdot V_1^+)^\omega, \mathit{Win})$ where $\mathit{Win} \subset (V_0^+ \cdot V_1^+)^\omega$ is Borel is determined.

Proof. We proceed in two steps. We first assume that the underlying vocabulary is the same set V. That is $V_i = V \times \{i\},\ i \in \{0,1\}$.

For a sequence $\rho \in (V_0^+ \cdot V_1^+)^\omega$, where $\rho = (u_0^0 u_0^1)(u_1^0 u_1^1)\ldots$ with all the $u_k^i \in V_i^+$. Define $e_V(\rho)$ as the sequence of $(V^+)^\omega$ obtained by simply projecting the turns of ρ into V^+ (but preserving the turn structure). That is, with the notations we introduced earlier, $e_V(\rho) = [\pi_\star(u_0^0)][\pi_\star(u_0^1)][\pi_\star(u_1^0)][\pi_\star(u_1^1)]\ldots$. Note that this is a sequence of **strings** on V, not a sequence of elements of V: the turn structure is still present and their is no implicit string concatenation between the bracketed elements. Notice also that because we still have the turn structure e_V is a bijection between $(V_0^+ \cdot V_1^+)^\omega$ and $(V^+)^\omega$. Let $\mathcal{G}' = \mathrm{GS}((V^+)^\omega, \pi_\star(\mathit{Win}))$ denote the GS game over vocabulary V^+ with winning condition $\pi_\star(\mathit{Win})$. By Martin's theorem, one of the players has a winning strategy in $\mathcal{G}'$. Turns are exactly the same in both games, and trivially, an infinite sequence $\rho_{\mathrm{GS}} \in (V^+)^\omega$ of turns is winning in $\mathcal{G}'$ iff its disjoint-vocabularies counterpart ${e_V}^{-1}(\rho_{\mathrm{GS}})$ is winning in $\mathcal{G}$. Therefore a strategy σ is winning in $\mathcal{G}$ iff it is winning in $\mathcal{G}'$. All that remains to show is that a set is Borel in $(V_0^+ \cdot V_1^+)^\omega$ iff it is Borel in $(V^+)^\omega$. It suffices to note that e_V is an homeomorphism between both the sets: e_V is clearly bijective, and the preimage of a basic open set $u_0 \cdot u_1 \ldots u_n (V^+)^\omega$ under $\pi_\star$ is, with a slight abuse of notation (leting for a string $u = (x_k)_{k<l}$, (u, i) denote the sequence of the $(x_k, i)_{k<l}$)

- if $n = 2k+1$

$$[(u_0,0)(u_1,1)]\ldots[(u_{2i},0)(u_{2i+1},1)]\ldots[(u_{n-1},0)(u_n,1)](V_0^+ \cdot V_1^+)^\omega$$

  and

- if $n = 2k$

$$\bigcup_{u \in V^+} [(u_0,0)(u_1,1)]\ldots[(u_{2i},0)(u_{2i+1},1)]\ldots[(u_{n-1},0)(u,1)](V_0^+ \cdot V_1^+)^\omega$$

In both cases the preimage is open so e_V is continuous. Conversely, ${e_V}^{-1}$ is continuous as well, since $e_V([(u_0,0)(u_1,1)]\ldots[(u_{2j})(u_{2j+1},i)](V_0^+ \cdot V_1^+)^\omega) = (u_0)(u_1)\ldots(u_{2j+1})(V^+)^\omega$. The Gale-Stewart determinacy theorem thus implies the Borel-determinacy of the ME game $\mathcal{G}$.

We now reduce the case of distinct vocabularies to the above. Assume now an ME-game $\mathcal{G} = \mathrm{ME}((V_0^+ \cdot V_1^+)^\omega, \mathit{Win})$ with *Win* borel and potentially $V_0 \neq V_1$. We will augment the vocabulary of both player with the moves of the other: Let $V = V_0 \cup V_1$, let $U_0 = V \times \{0\}$ and $U_1 = V \times \{1\}$. U_0 and U_1 are the 'augmented' vocabulary, we will reduce the initial game to a game played with vocabularies U_0 and U_1. Note that $(V_0^+ \cdot V_1^+)^\omega$ canonically injects into $(U_0^+ \cdot U_1^+)^\omega$. We first establish that if $W \subseteq (V_0^+ \cdot V_1^+)^\omega$ is Borel in $(V_0^+ \cdot V_1^+)^\omega$ it is Borel in $(U_0^+ \cdot U_1^+)^\omega$, which simply amounts to show that $(V_0^+ \cdot V_1^+)^\omega$ is borel in $(U_0^+ \cdot U_1^+)^\omega$. To get this, we note that

$$(V_0^+ \cdot V_1^+)^\omega = \bigcap_{n \in \omega} \bigcup_{u \in (V_0^+ \cdot V_1^+)^n} u \cdot (U_0^+ \cdot U_1^+)^\omega.$$

Since the vocabularies are countable, this set is indeed Borel in $(U_0^+ \cdot U_1^+)^\omega$. Consider now the set

$$S_1 = (U_0^+ \cdot V_1^+)^\omega.$$

It is easy to check thatd S_1 is also Borel in $(U_0^+ \cdot U_1^+)^\omega$. Let then $\mathcal{G}^0_\cup$ denote the ME-game over the common vocabulary U with winning condition $(Win \cup \overline{S_1})$. We have shown that this winning condition is borel in $(U_0^+ \cdot U_1^+)^\omega$. By reduction to the common vocabulary case, $\mathcal{G}^0_\cup$ is determined. In $\mathcal{G}$, any strategy for 0 picks, by definition, moves in V_0^+ only. Thus, to obtain a winning strategy for 1 in $\mathcal{G}$, it suffices to construct a strategy that wins and plays legal moves against any such strategy for 0, and similarly for 0, with V_1^+ replacing V_0^+. Now there are two subcases:

1. Assume that 0 has a winning strategy σ_0 in $\mathcal{G}^0_\cup$. Assume that 1 plays a strategy σ_1 for 1 **that only plays sequences in** V_1^+. Since σ_0 is winning against any strategy in $\mathcal{G}^0_\cup$, it is in particular winning against σ_1, and yield a sequence that is either in $\overline{S_1}$ or in Win. but against σ_1, this sequence might not be in $\overline{S_1}$ so it is in Win, and since by definition $Win \subseteq (V_0^+ \cdot V_1^+)^\omega$, this requires that, at any point, 0 plays in V_0. Hence σ_0 is a winning strategy for 0 in $\mathcal{G}$.

2. Assume 1 has a winning strategy in $\mathcal{G}^0_\cup$. Assume a strategy σ_0 that plays only sequences in V_0^+. σ_1 yields sequences in $(\overline{Win}) \cap S_1$. Hence yields sequences that are in S_1, which means that 1 only ever play in V_1, and that are $\overline{Win}$. Thus σ_1 a winning strategy for 1 in $\mathcal{G}$.

QED

Let us now go back to BM games and examine closely how 0-sum ME Games and a BM Game are related, depending on the winning condition:

7.3.3 Decomposition sensitive/invariant winning conditions

Assume that agents play with the same set of moves V *i.e.* $V_i = V \times \{i\}$. An important condition on Win is whether it hinges on which of the players made a particular discourse move. We call such winning conditions *decomposition sensitive*.

Definition 47 (Decomposition sensitive winning conditions)**.** $Win \subseteq (V_0^+ \cdot V_1^+)^\omega$ is *decomposition sensitive* iff $\exists W \subseteq \pi_{f,\omega}(Win)$ such that $\neg(\pi^{-1}_{f,\omega}(W) \subseteq Win)$.

Conversely, an ME game G with a *d*ecomposition invariant Win_G is one where: $\exists W \subseteq V^\omega$ such that a sequence $\sigma_1 \in (V_0 \cup V_1)^\omega$ is an element of Win_G iff $\pi_{f,\omega}(\sigma_1) \in W$. Thus, if 0 has a winning strategy in G, she also has one for attaining W in the BM game $BM(V, W)$, and conversely, if she has a winning strategy for attaining W in $BM(V, W)$, there is a sequence of plays that she can make regardless of what 1 does that will guarantee her a sequence $s \in W$. That sequence of plays yields a sequence $\sigma \in (V_0 \cup V_1)^\omega$ in which 0 and 1 are assigned different contributions at rounds such that $\pi_{f,\omega}(\sigma) = s$ and so $\sigma \in Win_G$. Thus, if 0 has a winning strategy in $BM(V, W)$, she also has a winning strategy in G. We have thus shown:

Proposition 2. Given an ME game $G = ((V_0^+ \cdot V_1^+)^\omega, Win_{ME})$ where Win_{ME} is decomposition invariant, 0 will have a winning strategy in G iff she has a winning strategy in the BM game $BM(V^\omega, Win_{BM})$ over V^ω where $Win_{BM} = \pi_{f,\omega}(Win_{ME})$.

In other words, when ME games involve decomposition invariant winning conditions, they collapse to BM games, and the existence of a winning strategy is predicted by the basic theorem for BM games which we expose now:

the Banach-Mazur theorem states necessary and sufficient conditions for the existence of a strategy to achieve Win. The theorem intuitively says that Player 0, the player who starts the conversation, can win

if her strategy takes into account, or has an 'answer', for almost all possible situations when her turn to speak may come. That is, the set of situations that her strategy doesn't take into account must be "small" in a topological sense:

Theorem 2 (Banach-Mazur (Mauldin, 1981))**.** Given a BM game $BM(V^\omega, Win)$, we have the following:

- Player 1 has a winning strategy if and only if *Win* is a meager set for the Cantor topology on V^ω;
- Player 0 has a winning strategy if and only if, there exists a finite string x such that $O(x) - Win$ is meager (that is, *Win* is co-meager in some basic open set) for the Cantor topology on V^ω.

Remarks It is worth noting the following:

- Since V_0 and V_1 are disjoint, the 'flattening' *flat* of a sequences of round $\rho \in (V_0^+ \cdot V_1^+)^\omega$ into a strictly alternating sequence $flat(\rho) \in (V_0 \cup V_1)^\omega \setminus ((V_0 \cup V_1)^* V_0^\omega \cup (V_0 \cup V_1)^* V_1^\omega)$ is easily shown to be an homeomorphism between the two sets of sequences equipped with their respective Cantor topology (or, more accurately for flattened sequences, the subspace topology on the set of strictly alternating sequences of the Cantor topology on $(V_0 \cup V_1)^\omega$). This shows, among other things, that the meagerness and the Borel complexity of W and $\text{flat}(W)$ are the same, so we might armlessly switch from "flat" sequences to ones structured in rounds and back again.
- In addition, we can show that the combinaison of flattening and projection $\pi_{f,\omega}$, of plays to sequences in V^ω, is a continuous function from $(V_0^+ \cdot V_1^+)^\omega$ to V^ω, it follows that the borel complexity of any set $U \subseteq V^\omega$, is the same as the Borel complexity of the set $\pi_{f,\omega}^{-1}(U)$ in $(V_0^+ \cdot V_1^+)^\omega$ but the converse is not true: it is **not** generally the case that the Borel complexity of W in $(V_0^+ \cdot V_1^+)^\omega$ is the same as that of $\pi_{f,\omega}(W)$ in V^ω. However, the converse is true for a decomposition invariant set $W \subseteq (V_0^+ \cdot V_1^+)^\omega$, since it that case we have $\pi_{f,\omega}^{-1}(\pi_{f,\omega}(W)) = W$. Hence in the case of decomposition invariant sets, and only it that case, we can indiferently study their Borel complexity or that of their projection on V^ω.

7.4 Constraints and the Jury

7.4.1 Concepts

While ME games as we have sketched them so far allow a player to say anything in a conversation, this is not conducive to most winning conditions. Certain general constraints are constitutive of winning conditions. Exploiting the resources of our discourse move vocabulary, we can define our constraints using the following concepts. We will need the following notations: let $\mathcal{T}$ be the set of rounds (recall that each round is a pair of moves of 0 followed by moves of 1) in an ME game and let $\mathsf{proj}_i : \mathcal{T} \to \wp(DU)$ be the projection from a round τ to the set of DU's labeling a move of the contribution by i (τ_i) therein. We naturally extend proj_i to finite or infinite sequences of turns ρ letting $\mathsf{proj}_i(\rho) = \bigcup_{k<\text{length}(\rho)} \mathsf{proj}_i(\rho_k)$. We will also assume feasible, given a finite sequences of turns ρ and labels $\pi, \pi' \in \mathsf{proj}_{i,j}(T)$ for some turn $T \in \rho$, to semantically define a predicate $attack(\pi, \pi', \rho)$ that is true whenever the conversational context constituted by the moves in the finite prefix ρ played so far is such that the move labeled by π' indeed represent an attack on the content of the move labeled by π. Defining and formalizing what exactly an attack is, is one of the main concern of chapter 8. Similarly, we will write $response(\pi, \pi', \rho)$ when ρ is such that π labels an attack on some other move, and that π' is a response to that attack (typically π' is itself an attack on π). Whenever the finite play ρ is clear from context we might simply write $attack(\pi, \pi')$ or $response(\pi, \pi')$.

- **Comparative NEC (CNEC)**: **CNEC** holds for Player $i \in \{0,1\}$ on round τ of a play ρ if there are fewer attacks on i with no response in $\mathsf{proj}_{1-i}(\tau')$ for $\tau' \leq \tau$ than for $1-i$, *i.e.*, defining $\rho_{<\tau}$ as the partial play up to τ, and the sets

$$A_i = \{\pi \in proj_i(\rho_{<\tau}) \mid \exists \pi' \in \mathsf{proj}_{1-i}(\rho_{<\tau})\ attack(\pi,\pi') \text{ and } \forall \pi'' \in \mathsf{proj}_{1-i}(\rho_{<\tau})\ \neg response(\pi'',\pi)\}$$

  we have $card(A_i) > card(A_{1-i})$. **CNEC** holds for Player $i \in \{0,1\}$ over a play ρ if in the limit there are more rounds of ρ where CNEC holds for i than there are rounds of ρ where CNEC holds for $1-i$ [we shall make this notion formal later, after we introduce the model of the jury].

- **Non vacuous consistency**: A play ρ of an ME game over vocabularies V_0 and V_1 is consistent for player i iff for any model $\mathfrak{A}$ such that $\mathfrak{A}$ updated by any initial segment x of ρ, $\mathfrak{A}^{\|x\|} \not\models C_i\bot$ and for some initial segment y of ρ, $\mathfrak{A}^{\|y\|} \models C_i\varphi$, where $\top \not\models \varphi$.

- **Coherence**: A contribution by Player $i \in \{0,1\}$ is coherent on round τ if, considering the set $\mathsf{proj}_i(\tau)$ of all label of moves by i in τ, there is a single "top" label π_{top}, *i.e.* for every other label $\pi \in \mathsf{proj}_i(\tau)$ there is is a label $\pi' \in (\mathsf{proj}_i(\tau') \cup \mathsf{proj}_{1-i}(\tau'))$, where τ' is τ or some previous round, such that either $\pi'' : R(\pi,\pi')$ or $\pi'' : R(\pi',\pi)$ is a move in τ_i for some label $\pi'' \in \mathsf{proj}_i(\tau)$ and relation symbol R. In simple terms, this requires that the contribution in τ links to a previous contribution by either player.

- **Responsiveness**: Player $i \in \{0,1\}$ is responsive on round τ if there exists $\pi, \pi'' \in \mathsf{proj}_i(\tau)$ such that there exists $\pi' \in (\mathsf{proj}_{1-i}(\tau'))$ where τ' is the previous round and for some R we have $\pi'' : R(\pi',\pi) \in \tau_i$.

We think that these principles are elementary, reasonable constraints on winning conversations, in particular consistency, coherence and responsiveness. We take consistency to be the fundamental constraint, because inconsistency at the level of commitments implies a commitment to anything including to the opponent's winning condition. The motivation behind CNEC is that there are many times where a player cannot prevent an attack, and yet she can still retain intuitively a strategy for achieving her objectives by counterattacking. In the literature on argumentation, it is often assumed that any attack on an attack renders it moot and ineffective (Dung, 1995) (exploiting theoretical results in the domain of abstract argumentation could maybe help refine the definition of CNEC). CNEC requires that a string is winning for 0 if intuitively 0 defends her commitments more successfully than 1 defends hers in the sense that 0 has more successful unrefuted attacks on 1 than vice versa. Note that this condition is beyond the expressive capacity of first order logic over linear orders; we need at least first order logic with quantifiers (Libkin, 2004). In fact, we need slightly more than this for the following reason: such a winning condition is impossible for 0 to attain at every stage in the game; on her turn 1 can always pile on the attacks that have yet to be answered by 0.

7.4.2 The Jury

The Jury of ME games enforces these constraints by integrating them together with a players' goals into what we call *the Jury winning condition*. The Jury will penalize contributions that are not coherent, and it will penalize a player that is not responsive on her turn. While being incoherent or unresponsive on a turn is not a game changer; being inconsistent is–inconsistency makes the player automatically lose. In addition, our Jury is sensitive to attacks that it deems successful; and it is sensitive to ones with no reply and thus validates CNEC. The following model of the Jury with two components makes these claims concrete. We wish to stress that the Jury does not alter the structure of the game, rather provides a model of what a *linguistically realistic* winning condition is. The Jury's scoring is only used in the course of defining the Jury winning condition. Formally,

Definition 48. The Jury is a tuple $\mathcal{J} = (\{\varphi_i\}, \{P_j\}, \{c_{i,j}\})_{j\in\mathbb{N}, i\in\{0,1\}}$ where

- $\varphi_i \subseteq (V_0^+ \cdot V_1^+)^\omega$
- $\varphi_0 = \overline{\varphi}_1$. φ_i is *the Jury persuasion rule* for each player i.
- For each j, $P_j : \{\mathsf{GOOD}_0, \mathsf{BAD}_0, \mathsf{GOOD}_1, \mathsf{BAD}_1\} \to [0,1]$ is a probability function with $P_j(\mathsf{GOOD}_i) = 1 - P_j(\mathsf{BAD}_i)$.
- For each i, j, $c_{i,j} \in [0,1]$.

Definition 49. Given two set of plays Win_i, $Win_j \subseteq (V_0^+ \cdot V_1^+)^\omega$, representing the *initial objectives* or *initial winning conditions* of player i, j respectively, we say that:

- The Jury $\mathcal{J}$ is *aligned with* i if $Win_i \subseteq \varphi_i$, it is *biased against* i otherwise (if $Win_i \setminus \varphi_i \neq \varnothing$).
- If the Jury is aligned with both i and j we say that it is *aligned* otherwise it is *misaligned*.

Notice that the Jury is misaligned iff it is biased against one of the player. From the requirement that $\varphi_i = \overline{\varphi}_j$, it is aligned iff $Win_i = \varphi_i = \overline{\varphi}_{1-i} = \overline{Win_{1-i}}$ (thus the initial objectives are 0 sum). For 0-sum initial objectives, the Jury is misaligned iff it is aligned with one of the player, and biased against the other.

We will also use the same terminology (biased, aligned) relative to the *initial* ME Game $\mathcal{G} = ((V_0^+ \cdot V_1^+)^\omega, Win_0, Win_1)$ and the Jury.

Given an ME games play, the Jury assigns a rating, $\|\tau_k\| \in \mathbb{R}$, to the contribution in round τ_k ($0 \leq k < |\tau|$) with the following constraints:

- $\mathsf{coh}_i(\tau_k) = \begin{cases} 0 & \text{if } k \bmod 2 = i \text{ and player } i \text{ fails to respect coherence in } \tau_k \\ 1 & \text{otherwise} \end{cases}$
- $\mathsf{res}_i(\tau_k) = \begin{cases} 0 & \text{if } k \bmod 2 = i \text{ and player } i \text{ is not responsive in } \tau_k \\ 1 & \text{otherwise} \end{cases}$
- If i is inconsistent by round τ_k of ρ, then $\mathsf{cons}_i(\tau_{k'}) = 0$ for all $k' \geq k$. Otherwise, $\mathsf{cons}_i(\tau_k) = 1$
- In addition the Jury also assigns a value $\mathsf{win}_i(\tau_k)$ to every round τ_k as follows. Suppose ρ_{k-1} is the play so far and $\rho_k = \rho_{k-1}\tau_k$. Then $\mathsf{win}_i(\tau_k) = 1$ if $(k \bmod 2) \neq i$ or $\mathcal{O}(\rho_k) \cap \varphi_i \neq \varnothing$. Otherwise $\mathsf{win}_i(\tau_k) = -1$. That is, $\mathsf{win}_i(\tau_k) = 1$ if τ_k advances i towards the Jury persuasion rule φ_i[30]; $\mathsf{win}_i(\tau_k) = -1$ if τ_k takes i further away from φ_i.

The Jury also maintains a probability distribution over types: BAD_i and GOOD_i modeling the gain or loss of credibility that i has faced so far. At each round we write this probability as P_k, and it is defined as follows:

- $P_0(\mathsf{GOOD}_i) = 1$ and,
- $P_k(\mathsf{BAD}_i) = 1 - P_k(\mathsf{GOOD}_i)$ and,

[30] We require for a move to advance i towards his winning condition that this move does not totally prevent him to reach that condition. This is an approximation that should ideally be refined into a scalar evaluation of what the play actually brings to i. This might be difficult to achieve for purely infinite conditions such that, coming back infinitely often to a given topic (as reaching the topic once does not really alters what remains to do). An interesting perpective towards a better approximation is that of ressorting to *mean-payoff* games (Zwick and Paterson, 1995). Mean-payoff games allow to combine a purely infinitary condition (technically a parity condition), with opposed objective of maximizing (for 0) or minimizing (for 1) the average of local scores associated with each moves.

- if i successfully attacks $(1-i)$ at round k, then:

$$\begin{array}{l} P_k(\mathsf{GOOD}_{1-i}) = P_{k-1}(\mathsf{GOOD}_{1-i}|\rho_k) = c_k P_{k-1}(\mathsf{GOOD}_{1-i}) \\ P_k(\mathsf{BAD}_i) = c_k.P_{k-1}(\mathsf{BAD}_i) \end{array}.$$

  Where $0 \leq c_k < 1$ is a constant representing the severity of punishment per single move of a player $1-i$ by the jury ($c_k = 2/3$ looks reasonable, but this might be refined, for instance in order to make c_k dependant of the 'badness' of the attack). If noone successfully attack her opponent then $P_{k+1} = P_k$.

These two ingredients contributes to a definition of the Jury's evaluation in the following way: $\|\tau_k\|$ of the k^{th} round's benefits to i is given as:

$$\|\tau_k\|_i = \begin{array}{l} \mathsf{coh}_i(\tau_k)\mathsf{res}_i(\tau_k)\mathsf{cons}_i(\tau_k)P_k(\mathsf{GOOD}_i)win_i(\tau_k) - \\ \mathsf{coh}_{1-i}(\tau_k)\mathsf{res}_{1-i}(\tau_k)\mathsf{cons}_{1-i}(\tau_k)P_k(\mathsf{GOOD}_{1-i})win_{1-i}(\tau_k) \end{array}$$

And i's score for a play ρ is given as

$$\|\rho\|_i^{\uparrow} = \liminf_{n\to\infty}\left[\frac{1}{n}\sum_{k=1,\tau_k\in\rho}^{n}\|\tau_k\|_i\right]$$

Then for a sequence ρ, the Jury assigns a win to Player i only if $\rho \in \varphi_i$ and $\|\rho\|_i^{\uparrow} > 0$. Such plays form the *Jury winning condition* for player i. It is easy to see that if a player's winning condition in an ME Game $\mathcal{G}$ is her Jury winning condition, then she can win only with plays respecting consistency. She must also respect CNEC, and must satisfy responsiveness and coherence 'most of the time' (unless the opponent is really badly attacked, and even in that case, she must always eventually return to coherent and responsive moves).

Remarks A few remarks are in order at this point. First on the Jury itself:

- Note that our assumption $\varphi_0 = \overline{\varphi}_1$ on the Jury persuasion rule means that the Jury assigns 'at most' one winner for a play. Thus even assuming that an ME game $\mathcal{G}$ with winning condition Win_0 and Win_1 is non 0 sum and the Jury is aligned with $\mathcal{G}$, the game $\mathcal{G}'$ for the Jury conditions might not have win-win plays, only lose-lose ones. The other situation where φ_0 and φ_1 may not be complementary might also be explored. In that case, the Jury might assign a win to both the players.

- With the same notations and assumptions, in the case of a biased Jury, that is when $\varphi_i \cap Win_{1-i} \neq \varnothing$ for some $i \in \{0,1\}$, it might be the case that the Jury assigns a win to Player i in $\mathcal{G}'$ even though she loses in the original ME game $\mathcal{G}$ and conversely.

- The above implies that we have different types of ME game-Jury pairs $(\mathcal{G}, \mathcal{J})$ depending on whether $\mathcal{G}$ is zero-sum or non zero-sum and whether the $\mathcal{J}$ is biased or unbiased. We shall explore these scenarios and the impact of unawereness of the Jury's bias in our section on *misdirection*, Section 7.6.1.

Then, given the kind of evaluation that the Jury is performing, the meaning of a move is largely fixed by its consistent and coherent uses in context, how it can be attacked and how it can be defended, amplified on and so on. ME games thus enforce an exogenously specified notion of meaning, specified by linguistic theory. This includes implicatures, which would seem to mark an important difference with the signaling models of section 6.4. In signaling models, implicatures arise as a byproduct of cooperativity in the game's equilibrium; our model takes them to be provided by linguistic theory, and then predicts agents' attitude

toward them in the conversation's continuations. We could thus think of a signaling model as one of implicature generation, while our model is one of implicature 'survival'.

We think however that the situation is more complicated for two reasons. First, on a commitment-based view such as ours, the constraints of consistency and coherence determine implicatures. For instance, Quayle's implicature *K* in example (6.3.5) that he was comparable to John Kennedy as a politician translates in our model into the fact that it is consistent and coherent both for Bentsen to commit that Quayle committed to *K* and for Quayle to commit that he did not commit to *K*. Now as noticed earlier, a decision to exploit or to deny an implicature brings with it a commitment that the linguistic premises (cooperativity, sincerity, competence…) of the implicature's derivation hold, or do not hold. This being understood, it does not matter whether one implements those premises within a logical theory, or within a signaling game's utility profile. But linguistic constraints like coherence and consistency do further work. Consider, example example (6.3.2) with a fully cooperative Bronston. Where does the implicature to the "No" answer comes from in the first place? This implicatures is fundamentally tied to coherence. Inferring "Yes" through Bronston indirect answer makes him less coherent than inferring "No", because the "No" allows for an implicit contrastive discourse relation ("No I did not. [But] the company had one" while the yes would require an explicit marker "the company had one **too**" to infer a relation (this is moreover actually confirmed by the natural prosody of Bronston's answer). Any model would have to rely one way or another on a pragmatic theory to explain utilies and model this assymetry.

Second, we do not think that signaling games are independent of an exogenous linguistic theory with regard to implicatures. One of the main concerns of the model in Franke (2009) is to bring conventional meaning back into signaling model, which is not innocuous by constraining the set of player types in the game. In order to capture implicatures properly, one needs to make conventional meaning a part of the signaling model. To this end Franke (2009) suggests that the set of types constitute a potential answer set to the question under discussion; hence, determining the game's context requires a pragmatic model as well.

In conclusion, both signaling games and ME game need to appeal to an exogenous theory. Signaling games give a nice implementation of the Gricean theory where linguistic considerations can often be "hidden" into the game context, whereas ME games allow a higher level form of quantification over those possible game contexts, which is crucial to account for the possibility of Gricean or non-Gricean speaker.

7.5 Winning conditions and their complexity

Let us recap what we have done so far:

- we have examined BM Games, their limitations, defined ME Games and sketched an SDRT-inspired vocabulary to apply these games to the analysis of conversations. We have shown that both kind of games are determined for winning sets Borel in their respective relevant topology, and charaterized when an ME game's condition is simple enough for the game to be emulated by a BM game and winning strategies being predictable by the BM theorem.

- Looking more closely to the modeling of conversation, we have define a model for the Jury, as enforcing particular kinds of winning conditions. Several constraints —consistency, coherence, NEC contributes to this condition. These constrain the meaning of the signals players use independently of the players' beliefs or preferences, in contrast to other game theoretic frameworks.

Nevertheless, our model so far has not looked closely as what constitutes what we might call the initial or 'situation-specific' component of winning conditions, nor to the general 'shapes' of conversations that depend on them and the Jury constraints. We now investigate these questions.

A related problem is that of the Borel complexity of the respective winning conditions. We will quickly state the Borel complexity of strong appromixation (like being *always* coherent) of the different linguistic constraints entering the Jury condition. Then we will turn to a more detailed examination of (non-Jury part of) the conversational objectives that players choose, discussing their complexity as well, and when can, whether there is a winning strategy in the game for achieving them.

7.5.1 Complexity of purely linguistic constraints

We have placed necessary constraints on winning conditions in the previous section. We now quickly state their Borel complexity and draw some conclusions from this. Let $\mathsf{NEC}, \mathsf{COH}, \mathsf{RESP}, \mathsf{CONS}, \mathsf{CNEC} \subset (V_0^+ \cdot V_1^+)^\omega$ be the sets of sequences in $(V_0^+ \cdot V_1^+)^\omega$ which satisfy resp. the NEC, coherence, responsiveness, consistency and the CNEC constraints. We have

Proposition 3. $\mathsf{NEC}, \mathsf{COH}, \mathsf{RESP}$ and CONS are Π^0_2 sets.

Proof. We give the arguments for RESP. Those for NEC, COH and CONS are exactly similar. RESP can be written as follows:

$$\mathsf{RESP} = \bigcap_{n>0} \{\rho \in (V_0^+ \cdot V_1^+)^\omega \mid \rho_n \textit{ is responsive in each turn}\}$$

where ρ_n is the length-n prefix of ρ. Now

$$\{\rho \in (V_0^+ \cdot V_1^+)^\omega \mid \rho_n \textit{ is responsive for each turn}\}$$

is an open set. Thus RESP, being the countable intersection of open sets, is Π^0_2. QED

We now turn to the complexity of CNEC. The set $\mathsf{CNEC} \subset (V_0^+ \cdot V_1^+)^\omega$ is defined as

$$\mathsf{CNEC} = \left\{\rho \in (V_0^+ \cdot V_1^+)^\omega \mid \liminf_{n\to\infty} \frac{\textit{good attacks by } 0 \textit{ in } \rho_n}{\textit{good attacks by } 1 \textit{ in } \rho_n} \geq 1\right\}$$

That is, Player 0 has the upper hand over her opponent more often. Now by the definition of lim inf we have the following sequence of equalities

$$\begin{aligned}\mathsf{CNEC} &= \bigcap_{N>0} \left\{\rho \in (V_0^+ \cdot V_1^+)^\omega \mid \exists m > 0,\ \forall n > m\ \frac{\textit{good attacks by } 0 \textit{ in } \rho_n}{\textit{good attacks by } 1 \textit{ in } \rho_n} > 1 - 1/N\right\} \\ &= \bigcap_{N>0} \bigcup_{m>0} \bigcap_{n\geq m} \left\{\rho \in (V_0^+ \cdot V_1^+)^\omega \mid \frac{\textit{good attacks by } 0 \textit{ in } \rho_n}{\textit{good attacks by } 1 \textit{ in } \rho_n} > 1 - 1/N\right\}\end{aligned}$$

Now note that the set

$$\bigcap_{n\geq m} \left\{\rho \in (V_0^+ \cdot V_1^+)^\omega \mid \frac{\textit{good attacks by } 0 \textit{ in } \rho_n}{\textit{good attacks by } 1 \textit{ in } \rho_n} > 1 - 1/N\right\}$$

is closed. So, CNEC is a Π^0_3 set. Using ideas from Chaterjee (2007) we can also show that CNEC cannot be expressed as a set of Borel complexity ≤ 2. That is, CNEC is Π^0_3 hard. We thus have

Proposition 4. CNEC is a Π^0_3 complete set

Propositions 3 and 4 establish that linguistically driven winning conditions fix in the most realistic case winning conditions at Π^0_3 since $\Pi^0_2 \subseteq \Pi^0_3$. We note that to win according to the Jury has a Π^0_3 complexity.

Now that we have settled the discussion for generic kind of linguistic constraint that the Jury enforces, let us examine in more details some different kind of, more situation-specific conditions that can be encountered and enter the player objectives:

7.5.2 Situation-specific conditions: reachability and safety

Let's start by looking at the winning condition for the conversation in example (6.3.1), which, at least on a certain interpretation, is a very simple, decomposition invariant condition. Suppose that in order to achieve her winning condition, it suffices for candidate A that she, or her interlocutor, mention at some point, it doesn't matter when, the theorem that she has proved (leaving aside the constraints NEC, NR, consistency or discourse coherence, which make winning conditions more complex). In other words, for A to win the game, her conversation x must eventually contain this move. More generally, conversations in which the objective of a player is to simply "touch upon" a certain topic exhibit the following shape:

$$Win = Reach(R)$$

Reachability, a characteristic property of many conditions classified as Σ^0_1 in the Borel hierarchy, is defined as follows. Given a non-empty subset $R \subset (V_0 \cup V_1)$ of the elements of the vocabulary, a string x in $(V_0 \cup V_1)^\omega$ is said to reach R if the elements from R occur somewhere in x. More formally, for a string x over the vocabulary $(V_0 \cup V_1)$ we let $x(i)$ denote the ith element of x. We define $occ(x) = \{a \in (V_0 \cup V_1) \mid \exists i, x(i) = a\}$ to be the set of all the elements of $(V_0 \cup V_1)$ which occur in x. For $x \in (V_0^+ \cdot V_1^+)^\omega$ we let $occ(x) = occ(flat(x))$. Then

$$Reach(R) = \{x \in (V_0^+ \cdot V_1^+)^\omega \mid R \subseteq occ(x)\}$$

is the set of all strings in which the elements of R occur at least once. The reachability set R of example 1 is the two elements set $\{(a, 0); (a, 1)\}$ with a the move of mentioning A's theorem. Since this winning condition is decomposition invariant, by the BM theorem A has a winning strategy, since the winning condition picks out an open set of strings of discourse moves.

Just as Reachability characterizes many conditions in Σ^0_1, *Safety* characterizes an important subset of these Π^0_1 conditions and is defined as follows. Suppose S (the 'safe' set) is a subset of $(V_0 \cup V_1)$.

$$Safe(S) = \{x \in (V_0^+ \cdot V_1^+)^\omega \mid occ(x) \subseteq S\}$$

is the set of all strings which contains elements from S alone. That is, the strings remain in the safe set and do not move out of it. One common sort of Π^0_1 condition is to prevent player 0 from reaching a Σ^0_1 condition; another is to avoid a certain commitment or finite set of commitments.

Note An alternative way of thinking about reachability and many Σ^0_1 conditions is to look at their definitions in terms of temporal logic. As our space of infinite strings is a set of linear orders, formulas of linear temporal logic (LTL) can describe some of its subsets, in particular many reachability conditions. For any element $a \in (V_0 \cup V_1)$, let the proposition p_a denote the property of visiting or playing a. For some finite $R \subset (V_0 \cup V_1)$, the LTL defining formula for the strings that reach R is:

$$\varphi_{reachable(R)} = \bigwedge_{a \in R} \Diamond p_a$$

where $\Diamond$ is interpreted as *eventually*.[31] A reachability formula of the form $\Diamond p_a$ is true at an index i of a sequence x ($x, i \vDash \Diamond p_a$) iff for some $j \geq i$, $x, j \vDash p_a$. A string x satisfies $\Diamond p_a$ ($x \vDash \Diamond p_a$) iff at the initial point 0 of x, $x, 0 \vDash p_a$. Safety also has an LTL defining formula for the strings that stay in S: where $\Box$ is interpreted as *always*,

$$\varphi_{safe(S)} = \Box \bigvee_{a \in S} p_a$$

For a safety goal of the form $\Box\varphi$, $x, i \vDash \Box\varphi$ iff for all $j \geq i$, $x, j \vDash \varphi$. $x \vDash \Box\varphi$ iff $x, 0 \vDash \Box\varphi$.

[31] For an introduction to LTL, see e.g. Lamport (1980).

The simple Σ^0_1 winning condition of candidate A's is decomposition invariant. However, this is not so for other goals. Consider example (6.3.3) again. Justin is playing a game with a disjunction of reachability conditions as his winning condition: his goal is to get Janet either (i) to admit that she has been seeing Valentino or admit that she hasn't. These characterize an open set in an ME game where 0 is Justin. Nevertheless, unlike candidate A's, Justin's winning condition depends on Janet's making a certain commitment, a decomposition sensitive winning condition. We need to analyze the topological characteristics of such winning conditions.

We first look at winning conditions that are decomposition sensitive in a particular way: *Win* depends on 0 making a particular contribution at each turn. We call these conditions *rhetorically decomposition sensitive*. Our constraints of responsiveness, coherence, consistency and NEC are all rhetorically decomposition sensitive conditions.

Proposition 5. If *Win* is rhetorically decomposition sensitive, then it is meager.

Proof. Let $x \in Win$ be a winning play. Since *Win* rhetorically is decomposition sensitive, for every prefix x_n of x which ends with a contribution of 1 there exists a finite y_n such that $\mathcal{O}(x_n y_n) \cap Win = \varnothing$. Since x was arbitrary, this means the closure of *Win* has empty interior. Hence, *Win* must be meager. QED

Given the constraints enforced by the Jury, we will be mostly interested in rhetorically decomposition sensitive winning conditions in ME games. Any winning condition incorporating these constraints is a meager set. However, not all winning conditions that are meager provide a winning strategy for Player 1.

Consider the following abstract ME game. Suppose $V = \{a, b\}$ and suppose Player 0 loses if and only if at any point she plays b. That is, the winning set *Win* is

$$Win = (V_0^+ \cdot V_1^+)^\omega \setminus Reach(\{(b, 0)\})$$

This is itself a rhetorically decomposition sensitive winning condition. Now, both *Win* and $\pi_{f,\omega}(Win)$ are meager sets in their respective topologies. As the Banach Mazur theorem rightly states, Player 1 has a winning strategy in the BM game $(V^\omega, \pi_{f,\omega}(Win))$: play b at some turn. However, she does not have a winning strategy in the ME game $(V_0 \cup V_1)^\omega, Win)$. That is because whatever she plays, Player 0 can always avoid playing b. In other words, the decomposition sensitivity of the ME games breaks down the applicability of the Banach Mazur theorem in ME games. Player 1 cannot 'play for' Player 0 now, which she can do in the BM game. A linguistic example of such a situation is a game G where Janet from example (6.3.3) is player 0. Janet has a winning strategy in G even though her winning condition is meager.

Conversely, consider a winning condition for 0 that depends on some finite number of contributions by 1. Call such goals *1-finite-decomposition sensitive*. An instance of such a winning condition would be Justin's. Recall that Justin's objective in example (6.3.3) is to get Janet to commit as to whether she has been seeing Valentino or not. Symbolize this commitment by Janet's as $(c, 1) \in V_1$, as Janet is Player 1. Then Justin's winning condition is the union of open sets $\{\mathcal{O}(x.(c,1)) : x \in (V_0 \cup V_1)^*\}$ and is co-meager. Nevertheless, his opponent Janet has a winning condition in such a game: never answer Justin's question directly. More generally,

Proposition 6. If an ME game G has a 1-finite-decomposition sensitive winning condition, then there is no winning strategy in G for 0.

Corollary 10. There are ME games with 1-finite-decomposition sensitive winning conditions that are co-meager, but where 0 has no winning strategy.

There are also 0-finite-decomposition sensitive winning conditions that depend only on finitely many moves of 0. These conditions are Σ^0_1 where 0 always has a winning strategy, as the BM theorem predicts.

A final situation is the one of the prosecutor in (8.2.4) has components that are decomposition sensitive, but the entire winning condition is not. The prosecutor's winning condition as described above is that Bronston must either commit to an answer or never answer *P*'s question. Given such a winning condition, there is a winning strategy for the prosecutor: keep asking the question until Bronston commits to an answer. In fact the entire game space is the winning condition for the prosecutor. Thus, we can infer:

Proposition 7. Decomposition sensitivity of winning conditions is not preserved under union.

Proof. Let Win_1 be a decomposition sensitive winning condition and let $Win_2 = (V_0^+ \cdot V_1^+)^\omega \setminus Win_1$. Clearly Win_2 is also decomposition sensitive. Indeed, because if the play is decomposed according to some play $u \in Win_1$ then Player 0 cannot win. However $Win_1 \cup Win_2 = (V_0^+ \cdot V_1^+)^\omega$ is clearly decomposition invariant. QED

We've now canvassed the whole spectrum of decomposition sensitive winning conditions in ME games. In general, decomposition sensitivity makes ME games more expressive and more complex than BM games, breaking the delicate link between topology and winning conditions given by the BM theorem.

Decomposition sensitivity also affects Borel complexity. If Player 0 has a winning strategy in a rhetorically decomposition sensitive ME game then the Borel complexity of *Win* is at least Σ_2^0.

Proposition 8. Let $G = ((V_0^+ \cdot V_1^+)^\omega, Win)$ be an ME game such that *Win* is rhetorically decomposition sensitive. If Player 0 has a winning strategy in G then the Borel complexiy of *Win* is at least Σ_2^0.

Proof. Let us mimic a winning play u of Player 0. $u \in Win$ only if for every prefix u_n, n odd, of u that ends in a Player 1 move, $\mathcal{O}(u_n) \cap Win \neq \varnothing$. Thus whatever 1 plays to reach u_n, 0 can choose v_n such that $u_{n+1} = u_n v_n$ is still a prefix of u. We may thus write u as

$$u = u_0\mathcal{O}(u_1) \cap u_2\mathcal{O}(u_3) \cap \dots$$

Hence *Win*, being at least a countable union of the above sequences, is at least Σ_2^0. QED

7.5.3 co-Büchi conditions

We've already met the next kind of set in the Borel Hierarchy, Σ_2^0 sets, also known as co-Büchi sets. Suppose C is a subset of $(V_0 \cup V_1)$ (the 'co-Büchi' set). Then

$$co\text{-}Büchi(C) = \{x \in (V_0^+ \cdot V_1^+)^\omega \mid \inf(x) \subseteq C\}$$

where $inf(x) = \{a \in V \mid \forall i, \exists j > i, x(j) = a\}$ is defined to be the set of all the elements of V which occur infinitely often in x.

In terms of LTL formulae, the co-Büchi condition may be viewed as follows. Let $C \subseteq V$ be the co-Büchi set. Then

$$\varphi_{co-Buchi(C)} = \Diamond \Box \bigvee_{a \in C} p_a$$

Classic examples of co-Büchi conditions are those with strings that eventually contain only elements of C or eventually settle down in C. That is, the strings eventually get stuck in the safe set C. example (6.3.6) is a motivating example for a conversation with a Σ_2^0 winning condition that also involves rhetorically decomposition sensitive conditions like responsiveness, coherence and consistency: Feynmann had to respond to his students' questions and in a coherent way lead them eventually to the topic that he wanted to discuss.

The winning condition that a conversation be finite is also a co-Büchi condition. A finite conversation is easily modeled in the ME framework; the initial segment in which the agreement is reached is then

succeeded by an infinite sequence of "null" moves that keep the content of the last move. Indeed, there is a close connection between agreement winning conditions and finiteness. If the only goal of the exchange is to achieve a fixed point in which the dialogue stays within this information state forever after, the conversation should stop once the terms of the exchange and the agreement are common knowledge. Being rational agents, our players will stop once they acquire the mutual knowledge that that state has been achieved and that nothing will take them out of it.[32]

Bargaining agreements or agreements on some permanent exchange of goods, which could also be information are naturally Σ^0_2 conversations even in the absence of these constraints—e.g., . Any information seeking conversation in which 0 has the goal of acquiring agreement about some intellectual issue φ, like a Socratic dialogue also has the structure of a Σ^0_2 winning condition. Walton (1984) calls these inquiry dialogues. Co-Büchi conditions distinguish between provisional and real agreement. In a provisional agreement, an agent provisionally may acknowledge another's contribution and agree to a bargain but later take the acknowledgment and the agreement back. If *Win* of the conversational game consists only in reaching a provisional agreement, *Win* is clearly Σ^0_1, as it does not constrain what happens after the provisional agreement is reached. Real agreement is different. Once attained between two agents, the agents do not deviate from it in any further conversation; no conversational moves take them out of that state of agreement, as required for a co-Büchi condition. This is essentially also a decomposition sensitive condition.

In the absence of any constraints, however, a decomposition sensitive winning condition of agreement for 0 has no winning strategy for 0; 1 always has a non empty 1-play of disagreeing for any possible continuation by 0. Similarly, no conversational goal of extracting a binding oath from an opponent can succeed, unless additional constraints are imposed. For similar reasons to the lack of a winning strategy for agreement type winning conditions, finiteness winning conditions are also easily seen to have no winning strategy—1 can always prolong the game by talking when it is her turn.

Our general constraints of NEC, responsiveness and coherence, however, can make the agreement happen in agreement seeking conversations. The Jury becomes an "arbitratror", imposing agreement when the opponent 1 no longer has any counter arguments to rebut 0's arguments for a particular position or exchange; the lack of counter arguments makes 1's objections not credible, thus lowering 1's score eventually leading to 0's winning condition. Thus, with the Jury's constraints, 0 wins a Σ^0_2 goal iff 1 has eventually no more arguments against a certain proposition φ, where φ may describe a bargain or topic of discussion.

Co-Büchi conditions also characterize goals in which 0 repeatedly attacks 1 eventually to reduce the opponent's score in the eyes of the Jury. Example (6.3.4) is such an example. Let LD be player 0 in an ME game. In the *voire-dire* transcript, 0 repeatedly returns to the question as to whether the defendant Tzeng was responsible for severing a nerve in a patient's hand; he seemed prepared to revisit the theme indefinitely until he exposes that the expert witness *D*, or 1 in this game, was covering up for a fault of the defendant. Repeatedly questioned, 1 replies each time in the play up to (6.3.4-c) to (6.3.4-d) that Tzeng was not at fault. In the Jury's eyes, 0's questioning had little effect; the Jury's probabilities assigned to the types of 0 and 1 did not shift, and 0 was no closer to his winning condition in getting the court to agree with him that 1 was not an impartial witness. However, at (6.3.4-d), 1 contradicts his previous testimony by agreeing to 0's loaded question, and his attempts to backtrack and correct his mistake are successfully attacked by 0 in (6.3.4-h). At this point, 0 has achieved his goal.

We note that our model of the Jury needs refinement in that it does not take account of successful retractions in the face of inconsistency, and so we cannot really predict the counterattack at (6.3.4-h). We plan to address this in future work.

[32] In principle, participants could continue acknowledging each other's acknowledgments *ad infinitem*. But such acknowledgments wouldn't serve any purpose. For a discussion see chapter 10.

7.5.4 Büchi conditions

The complementary condition of a Co-Büchi condition, the Büchi condition, is the equivalent of (infinite) iterated reachability, and is a condition that is not expressible on finite strings. Suppose B is a subset of $(V_0 \cup V_1)$ (the 'Büchi' set). Then

$$Büchi(B) = \{x \in (V_0^+ \cdot V_1^+)^\omega \mid \inf(x) \cap B \neq \varnothing\}$$

is the set of all strings which contain infinitely many elements of B or equivalently which visit B infinitely often. A Büchi set is Π_2^0 in the Borel hierarchy. In conversations where player 0 has a Büchi winning condition, she will win if she always has a path to B and revisits B infinitely often. A Büchi condition is more sophisticated than a Π_1^0 condition, in which a player never leaves a set of states. 0 can play for a Büchi condition and allow 1 a reachability or Σ_1^0 condition on his play. In such a game, player 0 can continue to return to her chosen and preferred states infinitely often, reiterating a point or set of points that she wants to make (once again, a finite conjunction of Büchi conditions is also Büchi).

Some Büchi conditions can be expressed using LTL formulas, however, that are finitely satisfiable. For any $x \in V$, let the proposition p_x denote the property of visiting or playing x. Let $B \subseteq V$ be the Büchi set of states. Then

$$\varphi_{Büchi(B)} = \Box \Diamond \bigvee_{x \in B} p_x$$

If 0's winning condition means revisiting a set of states B infinitely often, it must be for some other purpose other than agreements on exchanges of goods or information, for once lasting agreement is achieved, there is no point in revisiting that agreement. On the other hand, a Büchi condition can be effective in debate. Political debates like those evoked in example (6.3.7) exemplify a Büchi condition. Such a condition is more difficult to achieve if our rhetorical constraints are imposed on acceptable discourse sequences, because it means that any play by the opponent still must enable the player to have a rhetorically cooperative path to return to B. But a practiced debater can have such a strategy.

Let's now take a closer look at the analysis of one of our examples involving a Büchi winning condition, example (6.3.5). Our excerpted example was a turning point in the Vice-Presidential debate. Quayle's goal as Player 0 was to continually revisit the theme that despite his youth he had the talent and experience of a good Vice-Presidential and Presidential candidate. In effect this is a Π_2^0 winning condition. Up to the exchange in example (6.3.5), we can assume that Quayle had not made any disastrous moves, had remained consistent, responsive and replied to attacks and that the Jury's assignment to GOOD_0 had not suffered that much. His play had produced an initial segment of strings in *Win*. That is, we assume that the play σ up until example (6.3.5) is such that $\|\sigma\|$ was above 0, though not significantly above.

Given this goal, it would seem 0 had a clear winning strategy. What went wrong? To describe the exchange in example (6.3.5) in detail, we need as basic vocabulary for both V_0 and V_1: an attack move, $attack(x, y)$, meaning that the move y attacks move x, descriptions of the content of basic moves $x : \varphi$ and $y : \psi$, a commentary move, $comment(x, y)$ where a player expresses an opinion in y about move x, and a question answering move (QAP). (6.3.5-a) is a QAP move to a question about his Presidential qualifications. But the content of (6.3.5-a) is ambiguous. Quayle might have just intended (6.3.5-a)'s literal meaning—that he was equal in governmental experience to John Kennedy as a candidate for President. But he might also have intended, and probably did intend by mentioning a famous President, to have the audience and Jury draw a direct and positive comparison between himself and Kennedy with regards to the kind of President he might become in line with his winning condition. In response Bentsen plays $attack((6.3.5\text{-a}), (6.3.5\text{-b}))$. (6.3.5-b) introduces a commitment to a proposition φ, with φ contradicting the implicated direct comparison. This is what he should do; he should try to get the Jury to lower their estimation of Quayle's GOOD type. At this point Quayle should have counterattacked with another *attack* move, as NEC requires. But instead, Quayle plays a weak $comment((6.3.5\text{-b}), (6.3.5\text{-c}))$. (6.3.5-c) introduces

a commitment to a proposition ψ of Bentsen being "unfair" in the subsequent turn; and then Bentsen plays another successful *attack*(*comment*((6.3.5-b), (6.3.5-c)), (6.3.5-d)). (6.3.5-d) commits Bentsen to χ, where χ expresses that it was Quayle that brought up the comparison and thus opened himself up to attack—hence, *attack*((6.3.5-a), (6.3.5-b)) was perfectly fair. The Jury penalized Quayle severely for this double failure in replying to attacks, with a negative score for each of Quayle subsequent turns, and as a result, a failure to fulfill the Jury's condition which made Quayle lose the debate.

7.5.5 Muller conditions

A *Muller* condition is defined as follows. Suppose we are given a set $\mathcal{F} = \{F_1, F_2, \ldots, F_n\}$ of subsets of $(V_0 \cup V_1)$ (the Muller sets). Then $Muller(\mathcal{F}) = \{x \in V^\omega \mid inf(x) \in \mathcal{F}\}$ is the set of all strings which eventually (after a finite point) get stuck in one of the Muller sets, $Muller(\mathcal{F})$.

A Muller winning condition is a boolean combination of Büchi and Co-Büchi conditions. In terms of temporal logic formulae this can be seen as follows. Let $\mathcal{F} = \{F_1, F_2, \ldots, F_n\}$ be the set of Muller sets where each F_i is a subset of V. Then

$$\varphi_{Muller(\mathcal{F})} = (\varphi_{co\text{-}Büchi(F_1)} \vee \ldots \vee \varphi_{co\text{-}Büchi(F_n)}) \wedge (\varphi_{co\text{-}Büchi(F_1)} \Rightarrow$$

$$\bigwedge_{x \in F_1} \varphi_{Büchi(\{x\})}) \wedge \ldots \wedge (\varphi_{co\text{-}Büchi(F_n)} \Rightarrow \bigwedge_{x \in F_n} \varphi_{Büchi(\{x\})})$$

Since Muller conditions extend Büchi conditions, Muller conditions are not compatible with the goal of exchanging goods. Nevertheless, there are real life conversations with Muller "winning" conditions with multiple states, in which the participants revisit the states indefinitely often. In fact conversations with Muller winning conditions are commonplace. For instance, examples examples (6.3.5) and (6.3.7) have both Π^0_2 and also Σ^0_2 components to their winning conditions, as the Jury requires that they obey rhetorical cooperativity, NEC and consistency. Once a Π^0_2 winning condition is combined with a Σ^0_2 requirement, the result is a Muller winning condition.

Proposition 9. If 0 must obey a rhetorically decomposition sensitive condition with a Π^0_2 objective, then her winning condition is Muller.

There are also examples of conversations with Muller winning conditions. One involves a conversation between two partners who have lived for a long time together and who are quite old. After a certain point 0 always attempts to go through the same conversational moves, so that they can revisit the same memories, and touch on the same themes, laugh at the same jokes. 0 asks the same questions to get the same answers. To quote John Prine from the song *Far from Me*, *a question ain't really a question if you know the answer too.* In the song 1 plays along for a while, though she "waits a little too long," to laugh at the same, repeated jokes. In the end 1's goal is to break the cycle of repeated conversational moves by 0. Assuming that our conversational agents are rational, the goal of such a conversation is not information exchange or some sort of persuasion; it is something else like venting one's emotions albeit indirectly, reliving an experience, or conveying some other non literal message.

We have discussed an wide range of winning conditions, from the simplest to the more complex, in term of Borel complexity. All these conditions are in the Borel hierarchy, and though it is theoretically possible to define a non-Borel winning condition (at least up to the axiom of choice), it is not clear wether an actual player might adopt a condition as its conversational goal. Since we have, in earlier sections shown that for Borel conditions, 0-sum ME games are determined, this means that in the vast majority of instances of strategic conversations, the corresponding games are determined, which appeal to discuss the rational basis that player might have to start playing the game in the first place. The next sections adress this issue formalizing the notions of *misdirection* and *conversational blindness*.

7.6 Why talk?

In this section we propose an answer to the following : why would both players engage in a conversation if it is a zero-sum game?

Captain Kirk's response to an unwinnable test, the Kobayashi Maru Test, in the movie *Star Trek* was to change the conditions of the test. Our response is similar for a player involved in a zero sum ME game and whose opponent has a winning strategy: change the winning condition hoping to convince the Jury! The zero-sum ME-games framework is very simple, and only allows two parameters of variation for either player to achieve a win, when the opponent already has a winning strategy for achieving her winning condition:

1. choose a different winning goal for herself rather than the complement of her opponent's winning condition, or
2. change or expand her vocabulary

We explore these options in what follows.

7.6.1 Misdirection

Let us consider the first option. We have modeled conversations as ME games where the players have certain winning conditions Win_0 and Win_1. The ME game $\mathcal{G}$ might be zero-sum ($Win_0 = \overline{Win}_1$) or non zero-sum ($Win_0 \cap Win_1 \neq \varnothing$). We complexify the game a little, and assume that both the players are 'unaware' of the type of the Jury $\mathcal{J}$. Thus crucially, the players are unaware of the Jury winning conditions φ_0 and φ_1. Under such a setting 4 different situations might arise.

The simple cases are those where the Jury is unbiased, and $\mathcal{G}$ is either zero-sum or non zero-sum. In both these situations, the Jury will award a win to i if she meets her chosen winning condition Win_i and adheres to the Jury constrains of consistency, responsiveness and coherence.

The interesting cases arise when the Jury is biased. In that case, the Jury has a prior opinion of what 0 or 1 should say and forms its winning condition φ_0 and φ_1 based on these. A Jury biased against $1 - i$ allows i to 'misdirect'. That is, i plays so that even if she does not satisfy her winning condition Win_i in the ME game $\mathcal{G}$, she does (hope to) satisfy the Jury winning condition φ_i.

Since the Jury does not interact with 0 (or 1), our players must just take their best shot as to what the Jury's winning condition is. So 1 can have a winning strategy by transforming $\mathcal{G}$ by choosing a winning condition that is not the complement of 0's winning condition but that will nevertheless convince the Jury to award 1 the win in the evaluation of $\mathcal{G}$. This is what we call *misdirection*. Note that this renders $\mathcal{G}$ technically non zero-sum.

A similar argument holds for a non zero-sum ME game as well. Thus we can formalise misdirection as follows.

Definition 50. Let $(\mathcal{G}, \mathcal{J})$ be an ME game-Jury pair. $\mathcal{G}$ is prone to misdirection when

- $(\mathcal{G}, \mathcal{J})$ is misaligned, i.e., $\mathcal{J}$ is biased.
- There exists a play ρ of $\mathcal{G}$ and $i \in \{0, 1\}$ such that $\rho \in (\varphi_i \cap Win_{(1-i)})$.

We say that ρ is a misdirection of the ME game $\mathcal{G}$ and that Player i can misdirect if she has a strategy σ_i to achieve ρ.

Consider again example (6.3.2) involving Bronston (B) and the Prosecutor (P) which is plausibly a case where B and P are playing a zero-sum ME game with winning conditions Win_B and Win_P respectively. P is playing for the winning condition Win_P, in which B plausibly commits to

1. having had a Swiss bank account or,
2. or having never had one, or
3. refusing to answer the question.

It is clear that by just posing his question P has a winning strategy: either B must at least implicate an answer by linking his response to the question, or he refuses to answer the question in this round; if B does not answer, P can keep reposing his question, and thus establishes in the limit that B commits to refusing to answer the question. In fact Win_P so described encompasses all of the game space. Though ideally B would have preferred not to commit at all concerning bank accounts, had B taken his winning condition to be $\overline{Win_P}$, he would have lost. So it is rational for B to misdirect - choose a different goal in the hope that the Jury will be persuaded not to convict him in its evaluation of the ME game with respect to the goal of P's conviction. Given the goal set by P, B sets a goal of minimizing his losses and chooses a goal that may enable him to escape conviction.

What happens then is that B chooses a goal Win'_B that is the complement of a stronger winning condition than Win_P, the condition being that either B commits to a direct answer to (6.3.2-c) or to refusing to answer the question. This transforms the original zero-sum ME game to non zero-sum. He then fashions a winning strategy that satisfies both Win'_B and Win_P. Given that P has adopted Win_P as a conversational goal, he is rationally happy with any continuation of $\mathcal{O}(abc)$. So while P *would* have had a winning strategy for $\overline{Win}'_B$, by asserting (7.6.1), it is not necessary, given our characterization of Win_P, that P picks (7.6.1) over any other continuation of example (6.3.2), which makes Bronston move to attempt the misdirection, at least rationalizable.

(7.6.1)e'. Prosecutor: I'm not interested in whether the company had an account there. I am interested in whether you ever did. Please answer my question.

Thus, P won the original ME game and B won the evaluation of it by the Jury. The utility of B's decision, and his assignment of a particular type to the Jury, paid off in the end. In effect the ME game between *P* and *B* had *two* Juries deciding as to whether *B* avoided conviction or not. While the lower court, Jury 1, was convinced by P's arguments and B was convicted of perjury, a higher court overturned the verdict of perjury and the judgment of the lower court that Win_P was sufficient for conviction. The higher court claimed that P should have achieved $\overline{Win}'_B$. This, then is a successful use of misdirection by B.

The outcome of each of the 4 situations mentioned above can be concisely represented in tabular form as follows.

	zero-sum $\mathcal{G}$	non zero-sum $\mathcal{G}$
unbiased $\mathcal{J}$	not prone to misdirection	not prone to misdirection
biased $\mathcal{J}$	prone to misdirection	prone to misdirection

Remarks A further, arguably more interesting observation follows if we assume incomplete knowledge concerning *i*'s goals on the part of $1 - i$ and the Jury winning conditions on the part of both the players. To model these we need to introduce 'types' for the players and the Jury and prior probabilities about the beliefs that the players have about the types of their opponents and the Jury. The players might dynamically update their beliefs about these types as the game proceeds based on Bayesian update. To model this we need to enrich our model of ME games to the Bayesian games framework. We do this in an ensuing paper.

In this framework, continuing with our courtroom example, suppose that B does not know whether in fact P has adopted Win_P or $\overline{Win}'_B$. It is rational on B's part to attempt to answer indirectly; if P is happy with his indirect answer, then B has at least an even chance to achieve Win_B, which is better than refusing to answer and thus failing to achieve Win_B with certainty. If *P* is not happy with the answer, he can then

use the continuation in *abcde*. In this case B will fail to block either Win_P or $\overline{Win}'_B$. However, given that B assigns equal priors to both Win_P and $\overline{Win}'_B$, he has a good shot at blocking one.

There is also a moral in these extended Bayesian ME games for a player (Player 0 say). Player 0 may have a winning strategy for a winning condition that she thinks will convince the Jury, but an ideal winning condition for 0 for should preclude any set C for which she might assign a positive probability to 1's winning in the eyes of the Jury, given that 1 plays to stay within C. If that is right, then intuitively in example example (6.3.2), P picked the wrong winning condition; he should have played the stronger condition.

Our model fills some gaps in the analysis of Asher and Lascarides (2013), who analyze misdirections in terms of two finite games with different utilities and moves. They do not explain the utilities assigned to Bronston, whereas we explain the utilities for these local moves in terms of the global goals of the two players. Asher and Lascarides (2013) also requires that P not be aware of the fact that Bronston could have chosen a direct answer to (6.3.2-c) instead of the indirect answer that he gave. This seems to us unrealistic, and we do not do this.

7.6.2 Conversational blindness

The second option for 1 in the face of 0's having a winning strategy in $\mathcal{G}$ is to change the game by expanding her vocabulary. Thus, Player 1 is playing with a vocabulary W_0 but Player 0 assumes her to be playing with a vocabulary V_0 which is a strict subset of W_0. Thus Player 1 has moves available to her that are unknown to Player 0. Thus 1 has moves available to her of which 0 is not even aware, and we call this option *conversational blindness*. We will show that conversational blindness can make it such that 0's winning strategy in $\mathcal{G}$ no longer works in $\mathcal{G}'$ and 1 has a winning strategy in $\mathcal{G}'$ instead.

In the previous section, we showed that in misdirection, both the players are aware of each others moves but have different winning conditions. Here we have a different situation. If one player is playing with a subset of the other's vocabulary, then the epistemic situations of our two players is rather different. The player with the smaller vocabulary can only play against an opponent of whom she assumes common knowledge of his own game. The player with the larger vocabulary, however, if she is aware of the situation, will model her opponent's strategies and winning condition over the smaller game but herself strategize in her larger vocabulary. In such a situation, the player with the smaller vocabulary herself has a winning strategy for her conversation only to discover that the strategy does not work given the moves her opponent uses against her, moves of which she was unaware when devising her winning strategy. It thus can behoove a player to play in a larger vocabulary of moves than her opponent is aware of. On the other hand, playing in a smaller vocabulary simply limits the player's vocabulary and does not improve her chances of having a winning strategy in all but the most pathological cases. We shall elucidate this scenario revisiting example (6.3.5).

In Asher and Paul (2013), it was shown that for $X \subsetneq Y$, a Borel subset A of X^ω will have a higher Borel complexity in Y^ω under certain conditions. Thus, even if Player 0 has a winning strategy for a winning condition *Win*, she might cease to have a winning strategy if the vocabulary of the Player 1 changes to the bigger set Y. That is because the complexity of *Win* might increase in Y^ω. Making use of that result we can show the following for ME games. Let $V_1 \subsetneq W_1$.

Theorem 3 (Asher & Paul, transferred to ME games)**.** Let $\mathrm{ME}_1((V_0^+ \cdot V_1^+)^\omega, Win)$ and $\mathrm{ME}_2((V_0^+ \cdot W_1^+)^\omega, Win)$ be two ME games and let $Win \subset (V_0^+ \cdot V_1^+)^\omega$ be Borel. We have the following in the Borel hierarchy (summarised by Figure 7.2):

1. (a) If $Win \in \Sigma^0_1$ in the space $(V_0^+ \cdot V_1^+)^\omega$ then $Win \in \Sigma^0_2$ in the space $(V_0^+ \cdot W_1^+)^\omega$; (b) If $Win \in \Pi^0_1$ in the space $(V_0^+ \cdot V_1^+)^\omega$ then $Win \in \Pi^0_1$ in the space $(V_0^+ \cdot W_1^+)^\omega$.

2. For $1 < \alpha < \omega$ and α odd, (a) If $Win \in \Sigma^0_\alpha$ in the space $(V_0^+ \cdot V_1^+)^\omega$ then $Win \in \Sigma^0_{\alpha+1}$ in the space $(V_0^+ \cdot W_1^+)^\omega$; (b) If $Win \in \Pi^0_\alpha$ in the space $(V_0^+ \cdot V_1^+)^\omega$ then $Win \in \Pi^0_\alpha$ in the space $(V_0^+ \cdot W_1^+)^\omega$.

3. For $1 < \alpha < \omega$ and α even, (a) If $Win \in \Sigma^0_\alpha$ in the space $(V_0^+ \cdot V_1^+)^\omega$ then $Win \in \Sigma^0_\alpha$ in the space $(V_0^+ \cdot W_1^+)^\omega$; (b) If $Win \in \Pi^0_\alpha$ in the space $(V_0^+ \cdot V_1^+)^\omega$ then $Win \in \Pi^0_{\alpha+1}$ in the space $(V_0^+ \cdot W_1^+)^\omega$.

4. For $\alpha \geq \omega$, if $Win \in \Sigma^0_\alpha$ (resp. $Win \in \Pi^0_\alpha$) then $Win \in \Sigma^0_\alpha$ (resp. $Win \in \Pi^0_\alpha$) on going from the space $(V_0^+ \cdot V_1^+)^\omega$ to $(V_0^+ \cdot W_1^+)^\omega$. That is, the sets stabilise.

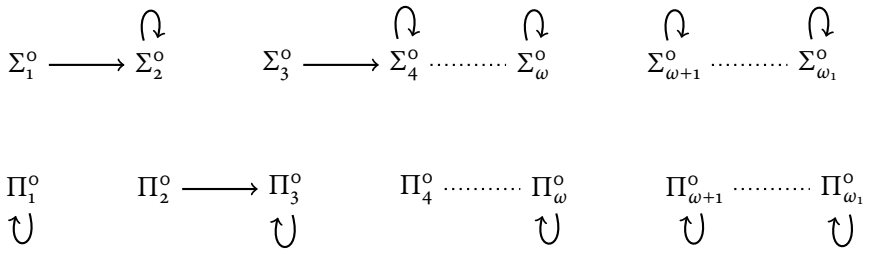


Figure 7.2: Jumps in the Borel hierarchy

The proof is general and applies to ω sequences over any countable vocabularies X and Y. We first prove the following:

Proposition 1. Let X and Y be two vocabularies such that $X \subsetneq Y$. An open set $\mathcal{O}$ in the space X^ω jumps to Σ^0_2 in the space Y^ω. A closed set C in the space X^ω remains closed in Y^ω.

Proof. In the following a subscript X or Y will denote the fact that we are refering to the space X^ω or Y^ω respectively. Let $\mathcal{O}$ be an open set in X^ω. Then $\mathcal{O}$ is of the form AX^ω where $A \subset X^*$. We first show how to code $\mathcal{O}$ as a Σ^0_2 set in Y^ω.

Each element u of A gives the basic open set $\mathcal{O}_X(u)$ which is a subset of X^ω. Now, when we move to the vocabulary Y, the set $\mathcal{O}_Y(u)$ is the set of strings which have u as a prefix and all possible continuations using letters of Y. Thus $\mathcal{O}_Y(u)$ is a strict superset of $\mathcal{O}_X(u)$. Hence, we need to restrict $\mathcal{O}_Y(u)$ in Y^ω such that we obtain a set which is equal to $\mathcal{O}_X(u)$ in X^ω. One way to do do so is as follows. Consider all the finite continuations of u in letters from X. For every $k \geq 1$, let $\mathcal{U}^k$ be the set of all the continuations of length $|u| + k$ and let $\mathcal{O}^k_Y(u)$ be the set

$$\mathcal{O}^k_Y(u) = \bigcup_{u' \in \mathcal{U}^k} \mathcal{O}_Y(u')$$

Note that each $\mathcal{O}^k_Y(u)$ is of the form AY^ω where A is a set of words of bounded length. Hence, each $\mathcal{O}^k_Y(u)$ is clopen. Thus, finally $\mathcal{O}_X(u)$ is the set

$$\mathcal{O}_X(u) = \bigcap_{k \geq 1} \mathcal{O}^k_Y(u) \tag{7.1}$$

which is a closed set, being a countable intersection of clopen sets.

Thus the set $\mathcal{O}$ can be represented in Y^ω as

$$\mathcal{O} = \bigcup_{u \in A} \mathcal{O}_X(u)$$

each of which by (7.1) is a closed set. Hence $\mathcal{O} \in \Sigma^0_2$ in the space Y^ω being a countable union of closed sets.

Let $\mathcal{O} \in (\Sigma^0_1 \setminus \Pi^0_1)$ in X^ω. To see that $\mathcal{O}$ is Σ^0_2 complete for Y^ω we first establish that $\mathcal{O}$ is not open in Y^ω. Indeed, because otherwise, there exists a finite string u whose all possible continuations with letters from Y are in $\mathcal{O}$ and that is a contradiction.

$\mathcal{O}$ is also not closed in Y^ω. To see this, we show that $(Y^\omega \setminus \mathcal{O})$ is not open in Y^ω. Since, $\mathcal{O} \in (\Sigma^0_1 \setminus \Pi^0_1)$ in X^ω, it follows from basic topology that $(X^\omega \setminus \mathcal{O})$ is non-empty. It also follows that there exists a string $u \in (X^\omega \setminus \mathcal{O})$ such that $\mathcal{O}_X(v) \nsubseteq (X^\omega \setminus \mathcal{O})$ for any $v \in \mathit{pref}(u)$, where $\mathit{pref}(u)$ is the set of prefixes of u. Let $u \in (X^\omega \setminus \mathcal{O})$ be such a string.

Suppose, for contradiction, that $\mathcal{O}$ is closed in Y^ω. Then $(Y^\omega \setminus \mathcal{O})$ must be open and hence must be of the form AY^ω for some $A \subset Y^*$. Now since $(Y^\omega \setminus \mathcal{O})$ should contain u there must exist $v \in A$ such that $u \in \mathcal{O}_Y(v)$. But then note that $\mathcal{O}_Y(v)$ also contains some string vu' such that $vu' \in (Y^\omega \setminus \mathcal{O})$ but which it should exclude. For example $vu' \in vX^\omega$ is such a string. This is a contradiction. Hence $\mathcal{O}$ cannot be closed in Y^ω.

Next let $\mathcal{C}$ be a closed set in X^ω. We claim $\mathcal{C}$ is also closed in Y^ω. To see this simply note that the complement of $\mathcal{C}$ in Y^ω can be represented as the open set

$$Y^\omega \setminus \mathcal{C} = \bigcup_{j \geq 1} (Y^j \setminus \mathit{pref}_j(\mathcal{C}))\, Y^\omega$$

where $\mathit{pref}_j(\mathcal{C})$ is the set of length-j prefixes of $\mathcal{C}$. Each $(Y^j \setminus \mathit{pref}_j(\mathcal{C}))\, Y^\omega$ is open. QED

Proof of theorem 3. Using the above lemma as a base case, an argument by induction proves theorem 2. QED

Theorem 3 shows that if a player i is playing with a particular winning condition *Win* that is Borel in a vocabulary V while her opponent is playing in a larger vocabulary W and she is unaware of or does not anticipate this fact, i may cease to have a winning strategy, because of the increased 'complexity' of the winning set *Win*.

Let's now return to example (6.3.5). It is a delicate matter to say exactly what Q did not anticipate and what went wrong. We give two interpretations; in both, Q plays a game with set of moves V and BN a game with set of moves W such that $V \subsetneq W$. One interpretation is that he did not foresee (6.3.5-b); alternatively, he foresaw (6.3.5-b) but did not believe that BN's move was relevant, because he failed to realize that (6.3.5-a) was ambiguous between the mere comparison of experience and the comparison between John Kennedy and Q as potential Presidents, the latter making BN's attack relevant.

We have seen that this goal would be characterized as Π^0_2 in the Borel Hierarchy in $(V_0^+ \cdot V_1^+)^\omega$. Q has a winning strategy for achieving this winning condition in ME_1. However, if BN can make a move not in V, then Theorem 3 says that *Win* jumps to Π^0_3 in $(V_0^+ \cdot W_1^+)^\omega$.

What does this mean for winning strategies in ME games? If the winning condition of an ME game is decomposition invariant, then one can apply the Banach Mazur theorem and infer that 1 has a winning strategy iff *Win* is a meager set. In ME_1, the game Quayle is aware of and is playing in, *Win* is Π^0_2, and it is a co-meager set. So by the Banach Mazur theorem, Quayle has a winning strategy in Π^0_2 in ME_1. But if *Win* is a Π^0_3 set in $(V_0^+ \cdot W_1^+)^\omega$, the encoding technique of theorem 2 makes this set meager in the larger topology and so Quayle ceases to have a winning strategy in ME_2.

The presence of a winning strategy is a complex matter in ME games, however, as we have established, because many winning conditions are not decomposition invariant; a winning condition may depend on finitely or infinitely many moves of one or both players.

Quayle's winning condition clearly depends on his contributions at every turn, and so it is decomposition sensitive. This means that a simple jump in Borel complexity or a move from a co meager to a meager winning condition is not enough to preclude a winning strategy.

Quayle's problem is that he does not refute Bentsen's attack or even really answer it, we assume, because he had not anticipated this move. But though not refuting an attack is costly with respect to our Jury winning condition, it is not a game losing condition. So what went wrong with Quayle?

Our intuition is that by not replying to Bentsen at all, Quayle made a game losing move. (6.3.5-b) to (6.3.5-d) in discourse theory terms is a `Commentary` on Bentsen's attack move. Commentaries implicitly carry with them a commitment on their author to the content they are commenting on. Now if the `Commentary`'s target is the *content* of what Bentsen said, then this is devastating for Quayle. Bentsen is implicating by saying that Quayle is no Kennedy, something stronger, that he is not of Presidential material. With `Commentary` on the *content*, Quayle then commits to that content. In so doing he commits to his not being of Presidential stature when precisely his winning condition was to constantly come back to that commitment and reaffirm it. His commitments are now inconsistent, and inconsistency is a game-losing property of conversations given our Jury winning condition. Moreover, this was an inconsistency involving an intrinsic property of the Π^0_2 part of Quayle's winning condition.

There is an alternative interpretation of the `Commentary` move by Q. The `Commentary` move is not about the content of BN's move but rather about the fact that BN made this move. This seems more plausible and it commits Q on the face of it only to the fact that Bentsen made a particular discourse move. But by not counterattacking, Q sends a message that is terrible for him. First, he commits that the attack is coherent and responsive. Second, by not replying he concedes and commits to the proposition that the content of BN's move *and its implicatures* are not attackable. That is, Q has no means to refute the content of the attack. But this in turn implies that he implicitly must commit to their content. Hence, his non-reply makes his commitments look inconsistent, and inconsistency is a game-losing property of conversations given our Jury winning condition. Moreover, this was an inconsistency involving an intrinsic property of the Π^0_2 part of Q's winning condition. So to have a hope of winning, Q should have replied somehow to the attack.

Let us now look more closely at the details. An 'irrelevant' move by BN will not make a difference to Q's strategy. However, if we suppose that BN's vocabulary W_1 contains a *relevant attack* on (6.3.5-a) and it is irrefutable, then indeed Q has no winning strategy in ME_2.

To flesh this out, we need to know what a relevant attack is.

Definition 51. An attack move α by i is relevant when it is a responsive move to some contribution of j's in a finite play ρ such that $\alpha \vDash C_i C_j \psi$ and $\rho \vDash C^* \neg C_j \neg \psi$; and $\alpha \vDash C_i \neg \psi$. ψ is called a presupposition of the attack.

BN's move (6.3.5-b) presupposes that Q's (6.3.5-a) suggested that the junior senator was comparing his potential Presidential stature to that of John Kennedy's, and it goes on to attack that supposed contribution. Intuitively (6.3.5-a) does have this content at least as an implicature on one possible reading. If that is right, then BN's attack was relevant, according to our definition.

We now investigate irrefutable attacks. An irrefutable attack $R(\alpha, \beta)$ by j on a move α by i is a relevant attack by j and i possesses no move $\gamma \in V_i^+$ that attacks j's move β that is potentially plausible to the Jury. That is, while V_i^+ always contains moves that attack the coherence or the responsiveness of any move of the opponent, these attack moves will not be plausible to the Jury, if the Jury can attach the move of the opponent to the discourse context to form a coherent SDRS. Since players have common knowledge of the principles of coherence and responsiveness and are committed to those principles, from this, we can conclude that given that there is no such attack move possible by i means that such moves are inconsistent with prior commitments.

Just because a player has no way to attack a move does not mean that she agrees with its content. But this disagreement is irrelevant, if she cannot defend herself. By not defending herself she concedes the point of the opponent's move (Lipman and Seppi, 1995). In our framework, if j does not attack i's move α with content $\neg\psi$, then the implicature is that j concedes the contents of $\neg\psi$ and so $C_j \neg \psi$. In addition,

since such implicatures are common knowledge, we have at least defeasibly that $C^*C_j\neg\psi$. On the other hand, since $\neg\psi$ was an attack move by i on j, j must have committed to ψ in some previous move in α. We must thus have $\alpha \vDash C_iC_j\psi$, and so the attack α also has the content that $C_j\psi$, which means that an attack without a response yields $C_j(\psi \wedge \neg\psi)$. With this analysis, we now see that Quayle indeed committed to an inconsistency, that he was both of Presidential material and not, and this is sufficient for us to predict that this move made him lose the debate.

In fact our analysis predicts that *any time* there is an irrefutable attack on player i, then i loses.

Proposition 10. Given an ME game $\mathcal{G} = ((V_0^+ \cdot V_1^+)^\omega, Win)$, if for any finite play $\rho \in (V_0^+ \cup V_1^+)^*$ such that $\rho.(V_0^+ \cdot V_1^+)^\omega \cap Win \neq \varnothing$, 1 has an irrefutable attack α, then 0 has no winning strategy in $\mathcal{G}$.

Note that an attack α is such that $\alpha \vDash C_i\neg\psi \wedge C_iC_j\psi$. Without a refutation of the attack, 0 in effect commits to the content of the attack which is inconsistent with her commitments in σ.

This is too strong a result. We have not considered the possibility that a player may *correct* his own prior commitments and change them. Quayle cannot correct his prior commitments that are inconsistent with what he tacitly commits to in not responding to Bentsen's attack, because those prior commitments define his debate goal. But in other cases, conversationalists can concede an opponent's point without automatically losing. We will come back to this point shortly.

When can a player respond to an unforseen attack in an adequate way? If one cannot rebut an attack, an alternative is to avoid it or render it moot. This is akin to what is known as an *undercutting defeater* of a proposition in epistemology (Pollock and Cruz, 1999). As a start to understanding this strategy, consider the following observation:

Proposition 11. Let

$$\mathrm{ME}_1((V_0^+ \cdot V_1^+)^\omega, Win) \text{ and } \mathrm{ME}_2((W_0^+ \cdot W_1^+)^\omega, Win)$$

be two message exchange games such that $V_0 \subset V_1$, $W_0 \subsetneq W_1$. Suppose Player 0 has a winning strategy σ_0 in ME_1 and suppose that Player 1 cannot attach any move from $W_1 \setminus V_1$ coherently to any play ρ that is consistent with σ_0. Then Player 0 also has a winning strategy in ME_2.

0 can thus have a winning strategy in the larger arena, even if she is unaware of the moves of the other player. To formalize this point, assume (without loss of generality) that an attack move α by i attaches to j's last turn in a finite play ρ. We then define the set of commitments of j after a finite play ρ in a given model as the set $T_{\rho,j} = \{\psi \mid \rho \vDash C^*C_j\psi\}$.

Definition 52. Two consistent theories T_1 and T_2 in a language L are potentially separable iff there is a formula φ of L such that $T_1 \cup \{\varphi\}$ is consistent and $T_2 \vDash \neg\varphi$. φ is then said to separate T_1 and T_2.

This definition reduces to separability, a standard notion in model theory when T is complete (Chang and Keisler, 1973). We are interested in theories $T_1 = T_{\rho,j}$ above and $T_2 = T_{\rho,j} \cup \{\psi\}$ for some ψ in a given model $\mathfrak{A}$ after a finite play ρ.

Definition 53. A set $S \subseteq (V_0^+ \cdot V_1^+)^\omega$ is closed under addition of φ, if appending φ in a consistent, coherent way to a finite prefix of any play $\rho \in S$ yields a string in S.

Definition 54. After a finite play ρ an attack α by i on j is inseparable from $T_{\rho,j}$ iff α has a presupposition ψ, and $T_{\rho,j}$ and $T_{\rho,j} \cup \{\psi\}$ are not potentially separable.

It is immediate from Proposition 10 that if Player 1 has an inseparable and irrefutable attack against opponent 0 that can be coherently and responsively integrated into the play, 1 has a winning strategy.

On the other hand if players can be playing in different games, we can show that 0 may have a winning strategy in some cases.

Proposition 12. Given ME games $\mathcal{G} = ((V_0^+ \cdot V_1^+)^\omega, Win)$ and $\mathcal{G}' = ((W_0^+ \cdot W_1^+)^\omega, Win)$ with $V_0 \subset V_1$ and $W_0 \subsetneq W_1$, if 0 has a strategy in $\mathcal{G}$ for playing $\rho \in Win$ and for each irrefutable attack $\alpha_i \in W_1$ by 1 on a move by 0 in some finite prefix ρ' of ρ, there is a formula φ_i that potentially separates α_i from $T_{\rho',j}$ and Win is closed under addition of φ_i, then 0 has a winning strategy in $\mathcal{G}'$.

Consider an attack α that responds to β in ρ', a finite prefix of ρ, where α has presupposition ψ. So $\alpha \vDash C_i C_j \psi$. But since α is potentially separable, there is a formula φ that 0 may append to β via a relation like `Background` or `Commentary`. Instead of playing β, 0 now plays $\beta \cdot \texttt{Background}(\beta, \varphi)$. Since $\beta \cdot \texttt{Background}(\beta, \varphi) \vDash C^* C_j \neg\psi$, α is not a consistent response to β. Thus, by adding the consistent, separating formula φ to σ, 0 blocks the attachment of the attack move α to β in σ. Then 1 cannot use the attack move on pain of inconsistency, which means that he would lose if he played α, and so that the conversation will remain in *Win*. 0 now does this for each attack α_i on her moves β_i.

So far we have shown the persistence of a winning strategy for 0, if she can prevent an attack from being made under the constraints of coherence and consistency or if she can convincingly refute the attack. But often a conversationalist can be faced with an irrefutable attack that does not destroy her; a debater, for instance can admit error and retract a commitment without thereby losing the debate. We could call such moves *self-corrections* and they are part and parcel of concession moves. What a self-correction by player i does is remove commitments that i previously made. One way of modeling such an to use the "erasure" operation discussed in Serre (2004). We plan to examine the situation of counterattacks in future work.

We can apply our observations to example (6.3.5). We mentioned two interpretations, one where Q is simply unaware of BN's attack move (in which case he could do nothing) and one where Q was unaware that his contribution was ambiguous—i.e., in his language V_0 his contribution committed him merely to a comparison about experience. BN's move is relevant in both cases, because in W_0 commits Q to the possibility that Q committed to a comparison of the two men, and that is what BN seized on. We have assumed that Q was unable to attack BN's move directly. But he could, however, have deflected the attack, had Q been aware of and willing to commit to the ambiguity of his contribution, by adding a particular rider, e.g., the one emphasized in (7.6.2) to his response. BN's move then would have been incoherent and irrelevant.

(7.6.2)a'. Q: [...] I have far more experience than many others ... *Although I would not presume to be the man or the towering political figure that John Kennedy turned out to be*, I have now as much experience in the Congress as Kennedy did when he sought the presidency...

This deflection rider, however, has a cost. Would it have kept Q's conversational contributions within the winning condition that he set for himself? We have supposed that *Win* should convey that he was Presidential material by comparing himself to John Kennedy. But how far did the comparison have to go for him to achieve his objectives? If (7.6.2) does not affect Q's attaining his objectives, then he has a winning strategy to stay in Win_G in $\mathcal{G}'$. While (7.6.2) entails that Q commits to not having the stature of the great John Kennedy, it brings the comparison to mind, in an almost classic example of *praeteritio*. On the other hand, if Q's objective was to commit that he was equal to John Kennedy in stature (an unwise thing to do), (7.6.2) would take Quayle out of his winning condition. In that case, there was nothing for Quayle to do.

7.7 Conclusions

In this chapter, we have defined ME games. ME games are played over two disjoint vocabularies V_0 and V_1 symbolizing the conversational moves available for player i and j, respectively. Moreover, ME games plays are infinite, conversational agents altern turns in which they play sequences of moves of their vocabulary

without stoping after any finite number of steps [33]. These two characteristic features are the key to address the challenges laid out in chapter 6, regarding strategic contexts: first, the game's model per se imposes only that (finite) conversations have the structure of a monoid; this is one of the fundamental assumption in the studies of formal languages, and this level of abstraction (the same that we adopted in chapter 4), leaves us free to define an appropriate vocabulary and intepretation function evaluating conversations into a different semantic space. In other word, one can rely on her favorite symbolic framework to define her own semantic constraints on successful plays. Second, the game abstracts 'completed' conversation as infinite sequences, limit objects of the aforementioned syntactic monoid. As a consequence, although each finite conversational prefix might (and following the way we formalized our Jury, will) be evaluated, a definitive winner might only be declared over the completed, and essentially infinite, conversation. Hence, assuming a semantic interpretation expressive enough, it becomes possible to define, on semantic grounds involving the different players' commitments at each finite prefix, their conversational goals and analyze whether a conversation achieves or fails to fulfill those goals. Better, we can distinguish on a semantic basis too, generic linguistic constraints (consistance, coherence, replying to attacks) from situation-specific constraints, such as, coming back repeateadly to a given topic, or touching upon a certain subject. The infiniteness of the game then ensures that in order to satisfy the generic constraints, the 'dilemma' that more classical settings face (cf. section 6.4), namely assuming either aligned payoffs, or else lack of communication of any sort, fades away as agents may not defect or opt out of the conversation without exposing themselves to subsequent attacks.

In the particular case of player using a common vocabulary V, *i.e.* when there are two injections π_0, π_1 from V into V_0 and V_1 respectively, we can define a notion of decomposition invariant winning condition (w.r.t. V, π_0, π_1) for which the game can be encoded as a Banach Mazur game, and winning strategies topologically characterized. We have shown however that most natural constraints on conversations do not fall into this category and have a particular and surprising complexity in the Borel Hierarchy. We have also proven that ME games are determined. This led us to analyze what conversationalists can do when faced with an opponent who has a winning strategy in an ME game. ME games predict that there are only two types of rational response to such a situation, each of which leads to an interesting linguistic phenomenon. The first was misdirection and the second conversational blindness. As far as we know, the notion of conversational blindness, though intuitive, has not been explored before in any systematic way; and our approach to misdirection, we believe, fills in some blanks in previous accounts. All in all, we find ME games to be a powerful tool in the analysis of conversation, one that we plan to use in future research.

In practice, the predictive power of the model crucially depends on the faculty of the choosen vocabulary and interpretation to successfuly implements the linguistic constraints contributing to form an agent's winning condition. Regarding the constraints we have identified, this means that the underlying language must express at least, commitments, coherence links, attacks, and be subject to a notion of a consistency. These are codependent and their interplay must also be laid out clearly. We have already sketched some of the underlying relations: from section 7.6.2 we know that attack moves presuppose certain commitments, in particular commitments about others' commitments. Through all our examples we also know that implicit, implicated or ambiguous contents of one's commitment licence attacks and rebutal as well. The logical language of SDRT is a good starting point to capture each of this notion and their relative dependencies. It furnishes the kind of concepts (dicourse labels, relations) and interpretation mechanism (dynamic semantics changing the commitment slates of agents) that we need. Yet we will see that there is a need to alter the theory in order to integrate such things as commitment about commitments, ambiguous commitments with a sound notion of consistency and logical consequence. This is the object of the next part of the present thesis.

[33] Which, as we have seen, does not mean that we cannot encode a common decision to end the conversation, only that a player can never, on its own, impose an end to the conversation.

Part III

Public commitments and winning conditions

In part I we have been concerned with the logical form of discourse: what it must be, and how existing representations differ. A fundamental necessary condition, central to all candidate meaning representations is that they express more than the sum of their constitutive units' litteral meaning: implicit coherence links and other form of implicated content add to the meaning of one's contributions, and must be represented as well. Part II, examining conversation as a rational process, raised additional complications: it is well known that an essential property of the aforementioned pragmatic inferences is their *defeasability*, meaning that they can be cancelled or altered by subsequent moves, and whether such inferences eventually end up in the common ground, or whether they do not, depends on agents preferences over the possible conversational outcomes. These considerations have lead to approaches deriving meaning from equilibria that agents reach given their respective preferences. Yet it appears that, in case of divergent goals, such approaches make 'unsafe' implicatures essentially like nothing, collapsing for instance (misleading) indirect answers and utterances of random facts, which, we have argued, is unintuitive.

The problem is therefore that, in order to formalize the rationality of using a particular move, we need to refine the link between the implicatures that a given move triggers, and the way they affect commitments of one using that move. To achieve that, we need logical forms and a semantic interpretation that give us a clear understanding of how the semantic evaluation at a given point will be transformed in subsequent evolutions of the conversation, which appeals to the classical view of dynamic semantics where the interpretation of a move expresses the way it transforms an input context, *i.e.*, a *context change potential*. Indeed chapter 4 and chapter 7 share a possible answer to both the problem of semantically and dynamically comparing conversations (*via* a continuation-based pseudometric) and that of formalizing one's conversational goals and evaluating the rationality of a move given these goals (*via* ME games): in both cases the idea is to look at the set of *consistent* continuations of one's contribution. Each continuation, might or might not retain implicated content, and raise other kinds of ambiguity. Looking at the set of all continuations avoids the need to make a final 'guess' regarding the presence or abscence of such implicated content as part of the meaning of one's contribution. Alltogether, what we need a move's interpretation to tell us is

- When that move can be consistently used.
- How it restrains the sequence of subsequent moves that can be consistently played after it.

Making such a view as precise as possible in an SDRT-inspired framework is the goal of the present part.

Chapter 8

Commitments, credibility, coherence and the Jury

Contents

8.1 Introduction

Part II made a case that deceptive moves such as misdirections or lies, are common in strategic contexts (*e.g.*, in trials or political debates), and should be analyzed in terms of agents' commitments, and preferences thereon, rather that beliefs or other private attitudes toward the content of the different dialog moves. It is also commitments, or lack thereof, that subsequent 'attack' moves target.

Independently of the context being cooperative or uncooperative, the concept of objective, public commitment (Hamblin, 1987) captures the immediate change of the world that the action of sending of a message brings about. Changes in beliefs, realization of intentions, are secondary effects, that must be derived from the content of commitments and additional assumptions about agents' cognitive states. Section 6.4 argued that, while in cooperative settings, Gricean theories might sucessfully model such a derivation and yield simple, elegant theories of meaning as belief change (implemented, for instance, in signaling games), the link gets broken as soon as cooperativity is dropped. Only sometimes, asking a question has indeed the purpose of raising some uncertainity, but always, it represents a commitment of the speaker to a request for

the addressee to commit over some answer. Typically in a political debate, an agent A might ask a question to another agent B, even though A knows the answer to the question. In such a case A is just seeking for B's commitment to an answer. If B complies and provides an answer, it can be in A's interest to further challenge this answer, even knowing it is correct.

Modeling strategic contexts therefore requires a semantic theory of the objective commitments that agents can force out of each other; on that basis the ME games of chapter 7 then propose a model of conversation as a rational process. Addressing the above requires us:

1. to dispose of a logical model of dialog moves and their interpretation as commitments.
2. to determine semantically what constitutes an attack and
3. to distinguish between attacks from a semantic perspective.

In the next section, section 8.2.1, we precise our definition of attacks and credibility, linking these to linguistic public commitments. In section section 8.2.2, we give some examples of attacks on credibility and/or theoretically challenging forms of commitments. Section 8.3 proposes a first semantic model capturing nested commitments in a simplified SDRT framework, out the analysis. Section 8.4 reviews the Jury's credibility scoring of section 7.4 (through constraint CNEC). Section 8.5 discusses related work, some limitations of the model that we will address in subsequent sections and concludes.

8.2 Attacks and commitments

8.2.1 Defining attacks from commitments

An *attack* can be thought of as exposing a deceitful intention. But determining intentions behind speech acts is a tricky business we will not be getting into. In line with the introductory discussion, we consider notions of credibility and attacks depending on overt and public linguistic commitments by speakers.

Using commitments, we now precise these. A dialogue agent i is not credible after move m iff

- It is shown for some φ that after performing m i has committed to φ and φ is absurd or clearly refutable (shown to be inconsistent with a prior claim of the agent or a background common assumption),
- and that it was plausibly in i's interest to ressort to m, would this situation have gone unnoticed.

An *attack* by player j on the credibility of i occurs iff j commits to the following: i has performed m, i is committed to φ, $\varphi \vDash \psi$, ψ is absurd or refutable, and it is in i's interest to use m. m *enables* an attack on credibility iff j can consistently attach a move m' to m that entails such an attack. In terms of discourse relations, an attack m' by j typically attaches to i's attacked move m with a `Correction` relation. As Lascarides and Asher (2009) advocates, if such an attack is left without a reply, it is defeasibly inferable that i agrees to the attack's content, and therefore that i is not credible in the above sense: this intuition underlies the constraint CNEC in ME games (cf. section 7.4).

Our notion of credibility differs considerably from that employed in the signaling games literature where credibility is defined in terms of beliefs, typically in equilibrium (see for intance Farrell, 1993b; Franke et al., 2009). Our notion of credibility is complementary: since, we claimed, it is too difficult to get a good grasp of private attitudes in strategic contexts, we rather define credibility in terms of commitments and attacks, agent's interests and logical consequence, none of which depends on how a message affects the agents' beliefs. Moreover, credibility in our sense is an evaluation of agents themselves, performed by a third party (the Jury, see section 7.4), rather than an agent's evaluation of the messages he receives.

Messages affect credibility by the attack they enable or carry: if an agent commits to a proposition that has disputable consequences, she leaves the door open to subsequent attacks.

To flesh out our picture of credibility and attacks more precisely, we need to explain our notions of consequence and interest or preference. We appeal to two distinct notions of consequence classical in discourse theories: ordinary logical consequence and defeasible consequence. We will assume that our agents are logically (though not factually) omniscient and so if i publicly commits to φ he also publicly commits to ψ if ψ is a logical consequence of φ (which we write $\varphi \vDash \psi$). Agents also commit to implicatures that are defeasible but what we shall term normal consequences that interlocutors would draw upon learning that i commits to φ. Finally, implicatures may be more tentative, as when i draws attention to an alternative to something to which he is explicitly committed. We'll assume that implicatures are modeled in a defeasible logic using a space of preferred models of the conversation. We also allow that some weak implicatures may exist only in some of the preferred models while stronger ones are true in all preferred models. We thus distinguish between the following three levels of commitment.

- **Non-defeasible commitment by i to φ**: φ is a logical consequence of every possible interpretation of i's contribution.
- **Implicit defeasible commitment by i to φ**: the "prefered" interpretations of i's contribution entail φ.
- **Weak implicit defeasible commitment by i to φ:** some interpretations of i's contribution imply φ. Section 4 will provide more formal definitions.

We take preferences to be tied to a conception of rationality, provided by the setting of ME games of chapter 7: let us recall here the ingredients of ME games that are needed to precise what "interest" means in our definition of attacks.

A conversational game involves two agents (or players) $0, 1$, and a third party, *the Jury*, who observes and judges (via a scoring function) but does not participate in linguistic exchanges. The player's objectives bear two components: players adopt a winning condition, which is a set of conversations such that an unbiased Jury will declare o sucessful iff o manages to play the game so as to realize one of the conversation of the winning set **and** respect some generic linguistic constraints. Hence, a first component (that we call the initial winning set) concerns specifics about what is sufficient to make the Jury decide in o's favor, given that the necessary conditions of the second component (that we call the constraints of the Jury, or linguistic constraints) are respected. Staying credible *i.e.*, avoiding attacks, enters the second condition. We further assume these two sets of constraints to be defined on the basis of the players' respective commitments in each of the (finite prefix of the) sequences they accept. The first component may therefore set some contextual parameters, which are common knowledge among the players. For instance, at court, the Jury should know that an "honnest" expert witness must not share an interest with the defendent lawyer. Conversations in which the witness commits to such a shared interest should therefore enter the defendant initial winning set. To give another example, the constraints of the Jury should encode that some facts are irrefutable and commonly known as such, by having the Jury considering everyone to be committed to these. Thus either player denying one of those facts will be deemed inconsistent and fail the Jury condition.

Using this general set up, we can now precise what we meant earlier by "it is in i's interest to commit to φ if φ is not attacked", in the definition of an attack. This means that, **assuming that the Jury's condition is automatically fulfilled**, i committing to φ advances her toward her goal. In other words, φ advances i toward her initial winning set Win_i[34].

[34]To make this notion more formal, we can for instance use a similar definition as we used for the score win_i assigned at each turn by the Jury in section 7.4: namely that i's move doesn't exclude his initial winning set from all possible continuations. Of course this is a crude an approximation, which should ideally be refined to a scalar notion; the more in i's interest a move appears to be, the stronger an attack on that move. Here as well, using *mean payoff* games is something to explore.

For instance, one way to analyze example (7.2.1), is that the Reporter's (R) refusal (7.2.1-c) of Sheehan's (S) answer (7.2.1-b), implicates an attack, where R commits to the following:

1. S's role as a spokesman comes with an non-deniable commitment $C_S\varphi$ of honnestly and cooperatively respond to R's questions (at least, unless they are obviously out of topic).
2. after (7.2.1-b), S has acknowledged R's question and is yet not committed to an answer, therefore $\neg\varphi$ obtains.
3. S's preference is to avoid committing to an answer to the question, and given this, (7.2.1-b) is in S's interest in the sense we explained.

Notice, that in our setting, committing about the opponent's preferences is non vacuous even if we assume these preferences to be common knowledge between the players, as knowledge does not entail commitment (moreover they need not be known to the Jury).

After this exchange, Sheehan fails to address the attack; worse, he justifies his preference to not answer with (7.2.1-d) ("We are not going to respond to unnamed sources on a blog"), and thereby agrees that he has such preference whereas, at the same time, he does not deny his commitment to transparency as a spokesman. This makes him at least defeasibly inconsistent.

Globally, players will prefer moves which make them look good in the eyes of the Jury and make the other look bad, or at least worse. An attack by i on a player j's credibility is a way to make j look less good, as the CNEC constraint of the Jury conditions ensures. Part of i's looking good is to not make mistakes, to not invite attacks on her credibility, but to make herself look good a player must provide positive reasons for the position she favors. *Mutatis mutandis* for the preferences of player j. More generally: (i) our players must play moves that make them look good; (ii) if player i is rational, she will prefer moves that make possible moves that j cannot attack; (iii) between 2 moves that make i look good but make possible attacks, she will prefer the one with the more indirect or weaker damaging context, since a more indirect damaging consequence is one that has a rebuttal move *that's not what I meant to say*.

8.2.2 Examples of complex commitment dynamics and attacks

In this section we discuss some linguistic examples featuring different sorts of commitments and attacks on credibility. These examples involve not only commitments to propositions expressed by assertoric clauses but also to propositions involving coherence relations that link clauses, sentences and larger units together into a coherent whole. That is, players commit to a particular content and to its relations with what has been said before. In so doing a player may also commit to contents proferred by his conversational partner as in Asher and Lascarides (2003).

The following two examples provides a simple illustration of what will constitute this section and subsequent ones main semantic concern: how coherence links, ambiguity, and commitment about commitments interract. Consider first a case in which speaker A takes C's initial moves to be ambiguous:

(8.2.1)a. C: N. isn't coming to the meeting. It's been cancelled.
b. A: Did you mean that N. isn't coming because the meeting's cancelled or that the meeting is cancelled as a result?
c. C: As a result.

A's clarification question in (8.2.1-b) presupposes that C's initial contribution was ambiguous between a `Result` and an `Explanation` move (DeVault and Stone, 2007; Purver, 2004). We take this to imply at least a weak implicature for both readings, either of which a conversational participant could have exploited. This is something we want to model, and we'll see in the next example how such implicatures are exploited by an interlocutor.

Now consider the following example:

(8.2.2)a. C: N. isn't coming to the meeting. It's been cancelled.
b. A: That's not why N. isn't coming. He's sick.
c. C: I didn't say that N. wasn't coming because the meeting was cancelled. The meeting is cancelled because N. isn't coming.

This example illustrates how commitments embed. In (8.2.2-b) A commits to the fact that C committed in (8.2.2-a) to providing an `Explanation` for why N isn't coming, even though (8.2.2-a) is ambiguous. Only such a commitment explains why A attacks that commitment in the way that he does by giving an alternative `Explanation`. But in fact, C takes that commitment by A to have misinterpreted him; C commits in (8.2.2-c) that he committed in (8.2.2-a) to offering a consequence or `Result` of N's not coming to the meeting.

Note that while A attacks a move of C's in example (8.2.2), he does not attack C's credibility in our sense. But neither does this example provide a case of misleading implicature.

However, example (6.3.5) from section 6.3.1, for instance, does. Implicatures play a key role in this example where Quayle argues, against the thesis that his little governmental experience would make him unsuitable for the presidency, that Kennedy before him, with as much experience as he have, was able to handle the presidency. But this answer to the question suggests an implicit comparison between the two politicians (both junior senators from a state, each with little governmental experience) and gives rise to the possibility of interpreting Quayle's move as a stronger commitment that he would likely be able to handle the presidency in the same way that John Kennedy handled his, which, if not challenged would serve Quayle's claim better. Bentsen seized upon this weak implicature of Quayle's contribution and refuted it, indirectly exposing to the audience the self-serving nature of the comparison.

Example (6.3.2) is yet another attested example, in which Bronston, unwilling to make an explicit commitment to having or not having had an account in a swiss bank, chooses instead to defeasibly commit to an answer with 6.3.2-d in an attempt to avoid further questioning (Asher and Lascarides, 2012). We duplicate this example below:

(8.2.3)a. Prosecutor: Do you have any bank accounts in Swiss banks, Mr. Bronston?
b. Bronston: No, sir.
c. Prosecutor: Have you ever?
d. Bronston: The company had an account there for about six months, in Zurich.

It is interesting to consider a continuation of this in which the prosecutor would indirectly attack the response (8.2.3-d).

(8.2.4)e. Prosecutor: I would like to know whether you personally ever had an account there?

If Bronston is forced under the threat of perjury to answer affirmatively, his response in (8.2.3-d) now looks pretty deceiving. The natural thought arises: Bronston was trying to deceive us into taking him as committed that he did not have an account. Though the prosecutor did not proceed as in (8.2.4), had he done so he would have successfully attacked Bronston's credibility.

Let us finally also recall example (6.3.4), introduced in section 6.3.1 as well, which features a quite challenging structure of commitments: the example is an excerpt from a *voir dire* examination in Friedman and Malone (2010), where, the plaintiff lawyer (LP) interogates an expert witness (D), whose role as an indendant witness, we take it, presuposes him to be committed to share no interest with the defendant. This example is challenging because the plaintiff lawyer efficiently plays simultaneously on two different lines of attack: he has been repeatedly coming back to questions about, on the one hand, the social relationship between the expert and the defendant, and second, the division of a nerve during a surgery, with a twofold

objective of getting the witness to characterize the surgical operation as mishandled, or else undermining the witness credibility as tainted with partiality.

(8.2.5)a. LP: And also, he put an electrical signal on that nerve, and it was dead. It didn't do anything down in the hand, it didn't make the hand twitch?
b. D: Correct.
c. LP: And we know in addition to that, that Dr. Tzeng tore apart this medial antebrachial cutaneous nerve?
d. D: Correct.
e. LD: Objection.
f. THE COURT: Overruled.
g. D: Correct. There was a division of that nerve. I'm not sure I would say "tore apart" would be the word that I would use.
h. LP: Oh, there you go. You're getting a hint from your lawyer over here, so do you want to retract what you're saying?

D's error was to agreed to LP's loaded question, in which LP 'hides' the proposition that is really at issue, namely that Dr. Tzeng was negligent, as a presupposition by embedding it under a factive verb. This makes it difficult for D to answer the question in a straightforward way without embedding that presupposition into his commitments. LP then successfully attacks D's credibility in (8.2.5-f) when D attempts to recover from this mistake with (8.2.5-e), by seizing on a weakly implicated discourse connection between (8.2.5-c) and 8.2.5-e, in SDRT terms, a relation of $\texttt{Result}^*$ (the commitment in (8.2.5-c) caused the commitment in (8.2.5-e)).

Overall, the examples suggest two general methods of deception: moves that implicate propositions that cannot be committed to explicitly for strategic reasons, and moves that trap agents into making commitments they should rationally refrain from, in prevision of attacking subsequent correcting moves for betraying reprehensible preferences.

Another feature of attacks is that generally they work gradually in damaging an opponent's credibility. Perhaps no one move succeeds on its own in convincing the jury that the opponent is duplicitous or incompetent; rather a series of moves gradually move a jury to a skeptical view of the opponent over the course of a conversation. The victory conditions for our players are to succeed in eventually moving the jury to a position in which the opponent is no longer credible.

8.3 Nesting commitments in SDRT

8.3.1 Ingredients

We need a semantic (dynamic) model in order to formalize an analyzis of our examples and attacks on credibility. We have already seen that we need to model as part of a speaker's contribution not only its compositional semantics but also its illocutionary effects, in particular the implicit discourse links between utterances, as these can trigger or convey attacks on credibility. We will therefore build on Lascarides and Asher (2009), as SDRT already offers a formal, logic-based approach of dialogue content (semantics and illocutionary effects).

Lascarides and Asher (2009) models the semantics of dialogue by assigning to each conversational agents a *commitment slate*. Each commitment slate contains a list of propositions that an agent is committed to, which involves rhetorical relations as well as elementary propositions. Lascarides and Asher (2009) models explicit and implicit agreements and denials of one agent about another agent's commitments. However, the analysis of credibility threats requires that we go a step further. Conversational agents

explicitly or implicitly refer to, and dispute, others' commitments. They attack their opponent's credibility by exposing inconsistencies in something they claim the opponent committed to or implicated, and defend against such attacks by denying a commitment to content that the opponent claims they committed to or implicated.

We thus need to represent the commitments of all speakers from their own and their interlocutors' points of view, as in Stone and Lascarides (2010). Moreover, we need to represent arbitrary nesting of commitments explicitly. Recall example 8.2.2. In (8.2.2-b) A corrects C's prior utterance, and thereby commits that C is committed to a false proposition p (N. is not coming because the meeting is cancelled). C rejects A's correction. But what C rejects is not the proposition that corrects p, but A's commitment that C committed to p. Therefore, C also commits that A commits that C commits that p. Further, we need to distinguish between weak and strong commitments: when an agent tries to misdirect another, he might for instance give a weak commitment the look of a stronger one. Thus our dialogue model will add three things to Lascarides and Asher (2009): explicit nested commitments, the commitments of each agent from every agents' point of view and explicit strong and weak commitments.

According to what we have laid out in chapter 7, conversations proceed as follows: speakers alternate turns, each performing a sequence of discourse moves. Because we are interested in commitments and attacks, we will not import the full machinery of SDRT here. We will symbolize clausal contents within a propositional language, but incorporate labels for speech acts and discourse relations so that we can roughly express discourse-structures following Asher and Lascarides (2003). Crucially, however, our language allows us to embed discourse structures under 3 modal operators C_i, S_i and N_i, for each agent i. A discourse move for an agent i is defined as a proposition labeled by a discourse label (also called *speech act identifier*). A proposition is either a base-level proposition, or a formula expressing a commitment over a discourse structure, *e.g.*, i commits that a label have some particular content, or a complex formula $R(\pi_1, \pi_2)$ where R is a coherence-relation symbol and π_1 and π_2 are discourse labels. A complex formula recursively involves previously introduced discourse labels.

The modalities make the language more expressive, since we can express commitments of different agents to different contents for a single speech-act. The formula $\pi : \gamma$ describes a set of possible worlds in which the content assigned to the discourse label π is γ. A formula $C_i\varphi$ denotes the set of worlds in which the assignment of contents to labels is such that it makes i commited to φ. $N_i\varphi$ is weaker and means that i defeasibly commits to the contents of the formula φ. Discourse structure and commitments interract: when $C_i(\pi : \gamma)$ holds, it must also hold that i commits that the speaker of π commits to γ. Put another way, the semantics will ensure $(C_i(\pi : \gamma)) \rightarrow C_i C_{\mathrm{spk}(\pi)}\gamma$. More generally, given the assignements to *discourse labels* in different possible worlds, we retrieve commitments over *informational content* by looking at the content assigned to the labels of a given speaker.

This sketches the general picture, which we need to refine slightly: the language we consider inherits SDRT distinction between veridical and non-veridical coherence relations (depending or whether or not their semantic consequences entail the content of their arguments, see section 2.3), and SDRT hierachical organization of labels, allowing complex constituents (also referred to as complex discourse units, CDUs, in SDRT terms). Thus, in order to deal with non veridical relations like `Goal`, `Conditional` or `Correction` which do not import all of the content of their arguments into their host complex constituent, we define an agent's commitments as provided by the contents of the *maximal* labels of that agent only. These are the labels which are not used as argument of a relation as part of the content of another, complex, label of the same speaker. Hence, a speaker is committed to a content φ in a world w iff w is such that the content associated to one of her maximal labels entails φ. We will finally make the simplifying assumption (in this section and in subsequent ones) that there is only one maximal label per speaker at a given point in time. This assumption is an harmless one: the different semantics we propose will be easy to adapt to keep track of more than one maximal label per speaker (notice that more than one maximal label requires one contribution to be incoherent). Alternatively one can add a default veridical cunjuntive relational predicate

into the language with no semantic effects, and not sufficient to ensure coherence. Let us now describe the language in full detail.

8.3.2 Syntax and semantics

Assume a set Φ_o of base-level propositions, a countably-infinite set of labels Π, a finite set of relation symbols $\mathcal{R}$ and a set of conversational agents I. In order to keep track of which agent x performs speech act π, we assume Π partitioned in $card(I)$ disjoint subsets $(\Pi_i)_{i\in I}$. We define $\mathrm{spk}(\pi)$ as the unique $i \in X$ such that $\pi \in \Pi_i$.

Definition 55 (Syntax of (linearized) SDRSs)**.** The sets of propositions, $\Gamma(\Phi_o)$, and discourse moves $\Delta(\Phi_o)$, are defined by the following BNF grammar:

$$\Gamma(\Phi_o) ::= \varphi \mid R(\pi_1, \pi_2) \mid C_i\Gamma \mid N_i\Gamma \mid S_i\Gamma \mid \neg\Gamma \mid \Gamma \wedge \Gamma \mid \Gamma \vee \Gamma \mid \Gamma > \Gamma \mid ?(\Gamma) \mid \Delta$$
$$\Delta(\Phi_o) ::= \pi : \Gamma.$$

Where the nonterminal symbols Γ and Δ are shorthand for $\Gamma(\Phi_o)$ and $\Delta(\Phi_o)$ respectively, the variables π_i and φ respectively range over Φ and Π, and i and R respectively range over I and $\mathcal{R}$. In addition, we define $\gamma_1 \to \gamma_2$ as shorthand for $\neg(\gamma_1 \wedge \neg\gamma_2)$.

Our language has a dynamic, possible worlds semantics: the interpretation of a formula is a context-change potential *i.e.*, a relation between, input and output *states*. We will take our states to encompass a world and an assigment function, as is classical in dynamic semantics but also a vector c of $card(I)$ labels denoting the maximal label for each agent, and an equivalence relation $\sim$ over worlds, to represent polar questions. States are thus tuples $(w, c, \sim, \sigma)$. We let $c(i)$ denote the i^{th} component of c, *i.e.*, the maximal label for agent i, and $C = \Pi_o \times \cdots \times \Pi_{card(I)}$ denotes the set of such different maximal labels vectors. We call a pair (w, c) a c-world, and a triple $(w, c, \sim)$ a p-world. A state is thus a (p-world, assignment) pair.

To develop a treatment of our examples, we will need a basic account of polar questions. We therefore adopt a simplistic version of Groenendijk and Stokhof (1984) and take questions to semantically denote an equivalence relation over worlds (propositions denote sets of possibilities which are partitioned into equivalence classes raised by questions). For intance, the question *whether p?* partitions a set of worlds in two, those worlds at which p on the one hand, and those at which $\neg p$ on the other.

An assignment assigns a propositional content to discourse label at a given world. Since our proposition are partitioned sets of worlds, an assignment is therefore a function that assigns to a label and a world a set of c-worlds and a partition thereof. $\sigma(w, \pi)$ is roughly the (partitioned) set of c-worlds in which the interpretation of π at world w is true.

The semantics is defined relative to a model with a (dynamic) selection function representing a conditional modal operator that we will use to represent the defeasible commitment and suggestion modalities, N_i and S_i). The selection function, in a given state, output a (semantic) proposition representing 'normal' consequences of a given (semantic) input proposition:

Definition 56 (Model, states and assignment)**.** Let $\mathcal{W}$ denote a set of possible worlds, let $\mathcal{W}_C = (\mathcal{W} \times C)$ denote the set of associated c-worlds, let $E \subseteq \wp(\mathcal{W}_C)^2$ denote the set of equivalence relations over $\mathcal{W}_C$ and $\mathcal{W}_P = \mathcal{W}_C \times E$ denote the set of p-worlds, *i.e.*, the set of tuples $\langle w, c, \sim\rangle$ with $\sim$ an equivalence relation over c-worlds. The set (static) of semantic propositions is the set of partionned set of c-worlds: $\mathcal{P} = \wp(\mathcal{W}_C) \times E$ (a semantic proposition in that sense is thus a pair made of the 'classical' semantic counterpart of a proposition, a set of worlds, and a set of polar questions represented by an equivalence relation). The set of assignment over $\mathcal{W}$ is the set of functions $\mathcal{A} = \mathcal{P}^{\Pi\times W}$ from label,world pairs to semantic propositions: $\sigma(w, \pi) = (W, \sim)$ with $W \subseteq \mathcal{W}_C$ a set of c-worlds, and $\sim$ an equivalence relation on W. The set of states

over W is the set $\mathcal{S} = \mathcal{W}_{\mathsf{P}} \times \mathcal{A}$. Finally, a dynamic proposition is a relation between states: the set of dynamic propositions is defined as $\mathcal{D} = \wp(\mathcal{S}^2)$.

A model $\mathcal{M}$ is a tuple $\langle \mathcal{W}, v, \star \rangle$, where $v : \Phi \mapsto \wp(W)$ is a valuation function evaluating at which worlds a base-level proposition $\varphi_0 \in \Phi_0$ is true, and $\star : \mathcal{S} \times \mathcal{D} \mapsto \mathcal{D}$ is a dynamic selection function.

Given a model $\mathcal{M}$, the function $[\![\cdot]\!]_{\mathcal{M}}$ maps each formula δ of the language to a binary relation $[\![\delta]\!]_{\mathcal{M}}$ over states. Discourse-level assertoric propositions in $\Gamma(\varphi)$ act as filters that let through only the worlds at which the proposition is true; Discourse moves in Δ filters the assignments.

Another bit of needed machinery is for interpreting discourse relations. In our semantics each relation affects the contents assigned to its terms. Veridical relations like `Explanation` or `Result` will simply update the contextually given values to its terms with the semantic effects of the relation on those terms Asher and Lascarides (2003). Non veridical relations like `Correction` or `Alternation` place constraints on the truth of the contents associated with the terms at worlds verifying the relation in question.

To ease notational clutter (especially regarding tuples) we will use 'bar' variables like $\bar{w}$ to range over c-worlds and p-worlds, and variables $s, t, u, \ldots z$ for states. $\bar{w}$ and s thus abreviates tuples, and we write $\bar{w}[k]$ or $s[k]$ to denote the kth component of $\bar{w}$ and s respectively (starting at $k = 0$); if $\bar{w}$ is a p-world, for instance, $\bar{w}[0]$ is a world, $\bar{w}[1]$ a vector of maximal labels, and $\bar{w}[2]$ an equivalence relation. Assume now a model $\mathcal{M} = \langle W, v, \star \rangle$, and let p denote a dynamic proposition (*i.e*, a relation between states). we use both infix notation $s \; p \; s'$ and prefix notation $(s, s') \in p$ to denote that p relates s and s' (and similarly for other binary relations). We will slightly abuse the set-theoretic notations such as $\subseteq$ and $\setminus$ to ease manipulation of semantic propositions: for a semantic proposition $P = (W, \sim)$ and set of c-worlds W', we let $W' \setminus P = (W' \setminus W, \sim)$ and $\overline{P} = W_{\mathsf{C}} \setminus P$. For two semantic propositions $P = (W, \sim)$, $P' = (W', \sim')$ we let $P \subseteq P'$ be a synonym for $W \subseteq W'$ and $\sim \subseteq \sim'$. Let $\|?p\|^{\sigma,\sim}_{\mathcal{M}}$ denote the equivalence relation over c-worlds defined as

$$(\bar{w}, \bar{w}') \in \|?p\|^{\sigma,\sim}_{\mathcal{M}} \text{ iff } (\exists \bar{u}\ (\bar{w}, \sim, \sigma)\ p\ (\bar{u}, \sim, \sigma)) \leftrightarrow \left(\exists \bar{u}'\ (\bar{w}', \sim, \sigma)\ p\ (\bar{u}', \sim, \sigma)\right)$$

i.e. the relation making two c-worlds equivalent if and only if they both validate p or they both fail p. Finally, given a dynamic proposition p we let $\|p\|^{\sim,\sigma} = (\{\bar{w} \in W_{\mathsf{C}} \mid (\bar{w}, \sim, \sigma)\ p\ (\bar{w}, \sim, \sigma)\}, \sim)$ the (static) semantic proposition corresponding to the c-worlds that p transforms, given an assignation and partition. Conversely given a static semantic proposition $P = (W, \sim)$ we let $test^{\sigma}(P) = \{\langle (\bar{w}, \sim, \sigma), (\bar{w}, \sim, \sigma) \rangle \mid (\bar{w}) \in W\}$ the dynamic filter version of P. This is all we need to define the semantics:

Definition 57 (Semantics)**.** Below is the semantics of discourse propositions:

- $(\bar{w}, \sigma)[\![\varphi]\!]_{\mathcal{M}}(\bar{w}', \sigma')$ iff $(\bar{w}, \sigma) = (\bar{w}', \sigma')$ and $\bar{w}[0] \in v(\varphi)$.
- $(\bar{w}, \sigma)[\![R(\pi_1, \pi_2)]\!]_{\mathcal{M}}(\bar{w}', \sigma')$ iff $(\bar{w}, \sigma) = (\bar{w}', \sigma')$ and $\bar{w} \in \|R(\pi_1, \pi_2)\|^{\sigma}_{\bar{w}[0]}$.
- $(\bar{w}, \sigma)[\![?(\delta)]\!]_{\mathcal{M}}(\bar{w}', \sigma')$ iff $(\bar{w}, \sigma) = (\bar{w}', \sigma')$ and $\bar{w}[2] \subseteq \|?(\delta)\|^{\sigma,\bar{w}[2]}_{\mathcal{M}}$.
- $(w, c, \sim, \sigma)\ [\![C_i\delta]\!]_{\mathcal{M}}\ (w', c', \sim', \sigma')$iff $(w, c, \sim, \sigma) = (w', c', \sim', \sigma')$ and letting $(W'', \sim'') = \sigma(c(i), w)$ we have: $\forall \bar{w}'' \in W'' \exists s'''\ (\bar{w}'', \sim'', \sigma)[\![\delta]\!]_{\mathcal{M}} s'''$.
- $(w, c, \sim, \sigma)\ [\![N_i\delta]\!]_{\mathcal{M}}\ (w', c', \sim, \sigma')$ iff $(w, c, \sim, \sigma) = (w', c', \sim', \sigma')$ and letting $P = \sigma(c(i), w)$ we have $\forall s, t \in \mathcal{S}\ s \star ((w, c, \sim, \sigma), test^{\sigma}(P))\ t \Rightarrow \exists u \in \mathcal{S}\ t[\![\delta]\!]_{\mathcal{M}} u$.
- $(\bar{w}, \sigma)\ [\![S_i\delta]\!]_{\mathcal{M}}\ (\bar{w}', \sigma')$ iff $(\bar{w}, \sigma)\ [\![\neg N_i \neg \delta]\!]_{\mathcal{M}}\ (\bar{w}', \sigma')$.
- $s\ [\![\neg\delta]\!]_{\mathcal{M}}\ s'$ iff $s = s'$ and not $\exists s''\ s\ [\![\delta]\!]_{\mathcal{M}}\ s''$.
- $s\ [\![\delta_1 \wedge \delta_2]\!]_{\mathcal{M}}\ s'$ iff $\exists s''\ s\ [\![\delta_1]\!]_{\mathcal{M}}\ s''$ and $s''\ [\![\delta_2]\!]_{\mathcal{M}}\ s'$.

- $s \, [\![\delta_1 \vee \delta_2]\!]_{\mathcal{M}} \, s'$ iff $s \, [\![\delta_1]\!]_{\mathcal{M}} \, s'$ or $s \, [\![\delta_2]\!]_{\mathcal{M}} \, s'$.

- $s \, [\![\delta_1 > \delta_2]\!]_{\mathcal{M}} \, s'$ iff $s = s'$ and $\forall t, u \in \mathcal{S} \; t \star (s, [\![\delta_1]\!]) \; u \Rightarrow \exists v \; u \; [\![\delta_2]\!] \; v$.

Discourse moves are interpreted as follows:

$$(w, c, \sim, \sigma) \, [\![\pi : \delta]\!]_{\mathcal{M}} \, (w', c', \sim', \sigma') \text{ iff } (w, \sim, \sigma) = (w', \sim', \sigma'), c'(x) = \begin{cases} \pi & \text{if } x = i \\ c(x) & \text{if } x \neq i \end{cases}.$$
and letting $(W_\pi, \sim_\pi) = \sigma(\pi, w)$ we have: $\forall \bar{w}'' \in W_\pi \; (\bar{w}'', \sim_\pi, \sigma) [\![\delta]\!]_{\mathcal{M}} (\bar{w}'', \sim_\pi, \sigma)$

Remarks

- As suitable, whether a question is true at a given state s does not depend on the actual world $s[0]$, only on the partition of worlds $s[2]$. Note that, what the set-inclusion condition $s[2] \subseteq \|?(\delta)\|^{\sigma, s[2]}_{\mathcal{M}}$ requires, is, back in relational terms, that if $\bar{w} \; s[2] \; \bar{w}'$ then it must be that both $(\bar{w}, \sigma[2], \sigma[3])$ and $(\bar{w}', \sigma[2], \sigma[3])$ are transformed by p or that they both are not. Hence it means that "whether p?" is indeed one of the question encoded by the equivalence relation $s[2]$.

- The semantics for the defeasible commitments $N_i\varphi$ requires that the (semantic) proposition that i's commited to so far (*i.e.*, $\sigma(c(i), s[0])$) normally entails φ in $s[0]$.

- The 'suggestion' modality $S_i\varphi$ is simply the dual of the defeasible commitment modality. It means that some normal worlds given i's commitments in the actual state are φ worlds which implements the weak implicit defasible commitment.

- $N_i\varphi \vDash S_i\varphi$. Assuming that the selection function satisfies an axiom of facticity (*e.g.*, $\star(s, p) \subseteq p$) gives in addition $C_i\varphi \vDash N_i\varphi$.

The semantics of relation must be defined on an individual basis, but we have made available the ingredients needed to that end; for instance an intuitive approach to `Explanation` (left-right veridical) and `QAP`(non left veridical) could be:

- Left (resp. right) veridical relations must verify $\|R(\pi_1, \pi_2)\|^\sigma_w \subseteq \{(w, c, \sim) \mid \text{letting } (W_x, \sim_x) = \sigma(\pi_x, w_x) \text{ for } x \in \{1, 2\}, (w, c) \in W_1 (\text{resp. } W_2) \text{ and } \sim = \sim_1 (\text{resp. } \sim_2)\}$.

- $\|\texttt{Explanation}(\pi_1, \pi_2)\|^\sigma_w \subseteq \{(w, c, \sigma) \mid \star(w, test^\sigma(\sigma(\pi_2, w))) \subseteq test^\sigma(\sigma(\pi_1, w))$ and $\star(w, test^\sigma(\sigma(\pi_2, w))) \subseteq test^\sigma(\sigma(\pi_1, w))\}$ (this is a 'pure semantic' version of the clause $K_{\pi_2} > K_{\pi_1} \wedge \neg K_{\pi_2} > \neg K_{\pi_1}$ of section 2.3)

- $\|\texttt{QAP}(\pi_1, \pi_2)\|^\sigma_w \subseteq \{(w, c, \sigma) \mid \text{letting } (W_x, \sim_x) = \sigma(\pi_x, w_x) \text{ for } x \in \{1, 2\}, (W_2/\sim_1) \subsetneq (W_1/\sim_1)\}$ that is, the proposition represented by π_2 admits strictly less equivalence classes for questions in π_1, than the proposition in π_1 does. Therefore, π_2 induces a strict refinement of the partition of worlds by questions in π_1, and hence π_2 must resolve one of π_1's questions (note that this semantics entails that π_1 has more than one equivalence class, and thus indeed encodes at least one question).

This gives us a picture of how speakers' commitments evolve throughout a conversation.

Table 8.1: Analyisis of example (8.2.2) following Lascarides and Asher (2009).

turn	C's SDRS	A's SDRS
(8.2.2-a)	$\pi_1 : \neg N$ $\pi_2 : \texttt{ccl\_meeting}$ $\pi_3 : \texttt{Result}(\pi_1, \pi_2)$	
(8.2.2-b)		$\pi_4 : \neg\texttt{Explanation}(\pi_1, \pi_2)$ $\pi_5 : \texttt{Correction}(\pi_3, \pi_4)$

8.3.3 Examples revisited

We start with example example (8.2.2). Lascarides and Asher (2009) would analyse the two first turns as in table table 8.1

This representation is problematic; it makes A committed to an absurdity, here is why: the semantic conditions of $\texttt{Correction}(\pi_3, \pi_4)$ require that the content of π_3 implies the negation of π_4, but $\texttt{Result}(\pi_1, \pi_2)$ does not imply $\texttt{Explanation}(\pi_1, \pi_2)$ (the two are even contradictory). Keeping with the same kind of tabular representation that Lascarides and Asher (2009) use would require to further divide each cell of the table above in two, introducing A' interpretation of $C's$ moves, and to repeat this subdivision process potentially infinitely many times to express arbitrary nestings as in table table 8.2. Our language provides

C's SDRS	A's SDRS	
$\pi_1 : \neg N$ $\pi_2 : \texttt{ccl\_meeting}$ $\pi_3 : \texttt{Result}(\pi_1, \pi_2)$		
	C	A
	$\pi_3 :$ $\texttt{Explanation}(\pi_1, \pi_2)$	$\pi_4 :$ $\neg\texttt{Explanation}(\pi_1, \pi_2)$ $\pi_5 :$ $\texttt{Correction}(\pi_3, \pi_4)$

Table 8.2: Adding nested commitments

a well-defined syntax and semantics to implement this changes, for instance example (8.2.2) in its entirety is analysed as:

$$\begin{aligned}
&\pi_1 : \neg N \wedge \pi_2 : \texttt{ccl_meeting} \wedge (\pi_3 : \texttt{Result}(\pi_1, \pi_2) \vee \pi_3 : \texttt{Explanation}(\pi_1, \pi_2)) \\
&\wedge \pi_4 : \neg\texttt{Explanation}(\pi_1, \pi_3) \wedge \pi_5 : (\pi_3 : \texttt{Explanation}(\pi_1, \pi_2) \wedge \pi_5 : \texttt{Correction}(\pi_3, \pi_4)) \\
&\wedge \pi_6 : \neg C_C(\texttt{Explanation}(\pi_1, \pi_2)) \\
&\wedge \pi_7 : (\pi_5 : C_C\texttt{Explanation}(\pi_1, \pi_2)) \wedge \texttt{Correction}(\pi_5, \pi_6)
\end{aligned}$$

with $spk(\pi_1 \mid \pi_2 \mid \pi_3 \mid \pi_4 \mid \pi_6 \mid \pi_7) = C$ and $\text{spk}(\pi_4 \mid \pi_5) = A$.

The keys to this example's representation are thus, first, the representation of C's ambiguous move using the non-determinism of dynamic semantics with $\pi_3 : \texttt{Result}(\pi_1, \pi_2) \vee \pi_3 : \texttt{Explanation}(\pi_1, \pi_2)$ and

second the possibility of C and A to select different branches of that commitment. Moreover, we can propose a systematic explanation of the way SDRSs' derivation process must be adapted to yield such a logical form. The crucial point is that correcting move like π_5 triggers presuppositions: here, a presupposition that *C*'s move π_3 commits him to the negation of π_4's content *i.e.*, $\texttt{Explanation}(\pi_1, \pi_2)$. This presupposition must be accommodated as part of th content of π_5. Similarly π_7 is a correcting move too, so the negation of the correctant π_6, namely $C_C(\texttt{Explanation}(\pi_1, \pi_2))$, has to be accommodated as well as part of the content of π_7.

In example (8.2.2), the ambiguity stayed unacknowledged. But in (8.2.1-b), *A* takes *C*'s commitments to involve two possibilities, and he commits to not know which *C* has in fact committed to. Thus, in example (8.2.1), we must have:

$$\begin{aligned}&\pi_1 : \neg N \wedge \pi_2 : \texttt{ccl\_meeting} \wedge (\pi_3 : \texttt{Result}(\pi_1, \pi_2) \vee \pi_3 : \texttt{Explanation}(\pi_1, \pi_2))\\ &\wedge \pi_4 :?(C_C\texttt{Explanation}(\pi_1, \pi_2))\\ &\wedge \pi_5 : ((C_C\texttt{Result}(\pi_1, \pi_2) \vee C_C\texttt{Explanation}(\pi_1, \pi_2)) \wedge \texttt{Clarification-Q}(\pi_3, \pi_4))\\ &\wedge \pi_6 : C_C\texttt{Result}(\pi_1, \pi_2) \wedge \pi_7 : \texttt{QAP}(\pi_5, \pi_6)\end{aligned}$$

with $spk(\pi_1 \mid \pi_2 \mid \pi_3 \mid \pi_4 \mid \pi_6 \mid \pi_7) = C$ and $\text{spk}(\pi_4 \mid \pi_5) = A$.

That is, A asks, with π_4, a question of whether C has commited to $\texttt{Explanation}(\pi_1, \pi_2)$ or not. This question must coherently attach to π_3, so we infer that the link represent a question of clarification over the content of π_3. Now, as `Corrections` do, such a link also comes with its presuppositions, in that case, a presupposition that the content of π_3 is ambiguous between the question's answer sets. Hence, the logical form accommodates the ambiguity of π_3 as part of π_5. Finally *C* answers that he commited to a `Result` relation.

In example (6.3.5), Bentsen (B) seizes on a weak implicature of Quayle's (Q). Q explicitly commits to a direct comparison between his experience in government and that of the young JFK, but B corrects a more general equivalence between the presidential promise of JFK and his own. Let us symbolize the former with the atomic proposition $\approx\texttt{JFK}$ and the latter with a second atomic proposition $=\texttt{JFK}$. Let us further assume that our modal conditional validates a specifity principle (or 'penguin principle'), *i.e.*, an axiom shema $\vDash (p > q \wedge p > r \wedge q > \neg r) \rightarrow (q > \neg p)$ (see Asher and Mao, 2000). If we assume now, as a non-logical axiom that *being of equivalent stature as JFK* is a normal consequence of *having as much experience as JFK*, *i.e.*, we restrict our attention to models that let $\approx\texttt{JFK} > =\texttt{JFK}$ transform any input state, then the following must hold: Let δ in our language, represent an input context, and let π_δ denote the maximal label for i in δ. Then we have: $\vDash (\delta \wedge \neg(\approx\texttt{JFK} > \neg\pi_\delta) \wedge \pi_\delta : \approx\texttt{JFK}) \rightarrow S_i(=\texttt{JFK})$[35]. This means that, unless Quayle commits to something that is not normal given Quayle *being of equivalent stature as JFK*, if Quayle adds $\approx\texttt{JFK}$ to his commitments, one can take him as suggesting *being of equal stature as JFK*.

With this in mind, we can represent *Q* and *B*'s turns ((6.3.5-a) and (6.3.5-b)) as:

$$\begin{aligned}&\cdots \wedge \pi : \approx\texttt{JFK}\\ &\pi'_1 : \neg=\texttt{JFK} \wedge \pi'_2 : ((\pi : =\texttt{JFK}) \wedge \texttt{Correction}(\pi, \pi'_1))\end{aligned}$$

with $\text{spk}(\pi) = Q$ and $\text{spk}(\pi'_1 \mid \pi'_2) = B$.

Once again, we take B's corrective move in π'_2 to trigger a presupposition that the corrected move π commits its speaker to the negation of the correctant π'_1 which is accommodated in π'_2. The success of this attack relies on a decision of the Jury: does B's move $\pi_2 : (\pi : \texttt{JFK})$ represent an admissible claim in the context of Q's having performed $\pi : \approx\texttt{JFK}$? Intuitively, yes, since, we have seen, Quayle's move implicates $S_Q(=\texttt{JFK})$. But it should also be consistant for Quayle to deny this claim (without explicit self-correction).

[35] There is a slight abuse here as we did not let single labels π be proposisitions of our language, but it is a simple add defining $(\bar{w}, \sigma)[\![\pi]\!](\bar{w}, \sigma')$ iff $(\bar{w}, \sigma)\ test^\sigma(\sigma(\pi, w))\ (\bar{w}', \sigma')$.

If Quayle opts instead for a veridical comment on π_1', he embarks the content of π_1' under his commitment (which is not really a claim serving his objective), and he fails to attack in return.

This example illustrates two needs: one is to revisit a little the Jury's scoring, in order to accurately account for the different strength of commitments and hence, of attacks. The second is to augment the model capacity of representing agent's claim about their own commitments or other commitments, with a more constrained approach to when it is possible to consistently make these claims, given what has been acknowledged by the different agents. In the next section we therefore briefly discuss how to adapt the Jury's behaviour, then we explain some of the model's limitations that will be overcome in subsequent sections.

8.4 Revisiting the Jury's evaluation

Returning to the ME games framework (chapter 7), speakers' winning condition involve conveying commitments that will make them look good in the eyes of the Jury; they also want to make an opponent look bad if possible by attacking her weak points. The model of the Jury (section 7.4) must ensure that the moves open to a participant that lead her to her winning condition may decrease or even vanish if her credibility is repeatedly attacked. In terms of commitments, a player has to weight whether a move that makes her look good or make the other look bad, but can be attacked in return, is worth using.

This weighting of pros and cons, is accounted for in ME games looking at the infinite sequence of attacks and rebutals that might ensue the use of the move. What has changed in this picture, in light of this chapter's consideration is,

- that attacks have different strength,
- that commitments have different strength too, and that they might be ambiguous and,
- that the strength of an attack, and the strength of the commitment it attacks should be related somehow.

In the same order of ideas, the Jury's evaluation should also reflect the duplicitous nature of weak implicatures that agents don't dare put out as full commitments. So the update of the probability on $P_{k+1}(\mathsf{GOOD}_i) = c_k P_{k+1}(\mathsf{BAD}_i)$ should no longer hold c_k a constant, and rather depend on the strength of the commitment under attack: for instance, the weaker commitment, the more duplicitous the temptative, the stronger the attack.

But then, as another consequence, ambiguities and defeasible commitments make possible mutually attacking moves. Attacks on weak implicatures have a rebuttal: you misinterpreted what I said (not that this might renders the move useless for the player, as if *e.g.* Bronston in (8.2.4) explicitly denies committing to an answer). We might now reach situations in which, consistantly: i commits that he commited to $\varphi \wedge \neg\psi$, j commits that i commited to $\varphi \wedge \psi$, on the basis of which j issues an attack a on ψ and i a rebutal r ("I never said that"). Such a rebutal is a statement that j is wrong in her interpretation, it is not a proof that she is. Thus, unless j accepts i's correction, a satisfies our definition of an attack on r as well. In such a situation a attacks r and conversely.

To deal with mutually attacking moves, the Jury should evaluate CNEC in a branching way, along two different paths, considering one and only one of the two mutual attacks successful in each path (*e.g.*, one with a successfully attacking r and not conversely, and the other with r successfully attacking a and not conversely in the example above). The final score should be the average of the score on each path ponderated by the respective credibility score $P_k(\mathsf{GOOD}_i)$ for the path where i's attack is kept as winning over $1-i$'s one, and $P_k(\mathsf{GOOD}_{1-i})$ for the other.

With the above modifications, the upshot of the model then should be that agents pay dearly if their credibility is successfully attacked when they advance a weak implicature, as evidenced in example (6.3.5). This is because, accordingly to our first remark, considering Quayle's possible rebuttal ("I did not say that") as winning, impacts less Bensen's credibility than considering B's attack as successful over the rebuttal, since it attacks a duplicitous implicated content. So accordingly to our second change, evaluating the play with a ponderated average will tilt the balance towards Bensen.

This changes could also explain the assymetry between replying to an attack by denying commited to an attacked content, and preventing the attack using a deflecting rider beforehand (*e.g.* in example (6.3.5) the one we considered in section 7.6.2), making the second generally preferable, at least if one is aware of the attack and as no other rebuttal.

8.5 Conclusions

8.5.1 Related work

Our work assumes a commitment based view of conversation rather than one based on the internal, mental states of the participants Hamblin (1987); Traum and Allen (1994); Traum (2008) and builds on and complements the model proposed in Stone and Lascarides (2010), which in turn extends DeVault and Stone (2007). They introduce a dynamic Bayesian model for discourse actions based on prior moves. The work we present here is more limited in scope but also goes into more detail: the model details how attacks on credibility function with respect to various types of commitments that come from different kinds of discourse moves; we show that even in simple conversations levels of embedded commitments can be very complex (contrary to a suggestion of Stone and Lascarides (2010)); and The Jury's Bayesian update on player types details a part of the picture of Stone and Lascarides (2010).

Related to our work are also recent attempts to investigate argumentation in actual dialogue Cabrio et al. (2013). Argumentation theory provides a framework for analyzing attacks and counterattacks Dung (1995). We have given more linguistic detail on how such attacks are carried out and how this can affect ones' strategy in conversation. On the other hand, we have presented a general model for credibility in strategic conversation. Different contexts may affect the parameters of the model that we have set up. For instance, sometimes the Jury may be a participant in the conversation in the sense that it is allowed to ask questions, sometimes not. Given a particular context, Jury might also function according to persuasion rules that are different from the one considered so far. We have chosen simple settings to illustrate our model. Finally, we have not gone into the details of how particular conversational contexts may dictate specific linguistic forms of attacks and defense, e.g., Drew (1992). Overall the model is general enough, we believe, so that we can tune the parameters to fit the particularities of specific contexts.

8.5.2 Limits of the commitment logic

The dialog model we proposed suffers several expressive limitations that we will address in subsequent sections. Each follows from the interpretation of discourse moves we have adopted. Let us recall this interpretation:

$$(w, c, \sim, \sigma)\ [\![\pi : \delta]\!]_{\mathcal{M}}\ (w', c', \sim', \sigma') \text{ iff } (w, \sim, \sigma) = (w', \sim', \sigma'),\ c'(x) = \begin{cases} \pi & \text{if } x = i \\ c(x) & \text{if } x \neq i \end{cases}.$$

and letting $(W_\pi, \sim_\pi) = \sigma(\pi, w)$ we have: $\forall \bar{w}'' \in W_\pi\ (\bar{w}'', \sim_\pi, \sigma)[\![\delta]\!]_{\mathcal{M}}(\bar{w}'', \sim_\pi, \sigma)$

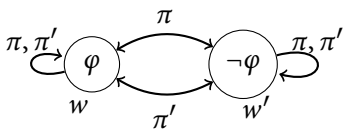
What is problematic, is that this clause makes our semantics dynamic in a very weak sense: given a sequence δ of discourse moves, a set of worlds $\mathcal{W}$ and a model over that set of worlds, each finite prefix

δ_i of δ of length i corresponds to a set of states S_i such that each set $s \in S_I$ is transformed by δ_i. However, the language offers absolutely no way to talk about which $s_i \in S_i$ 'survives' the $i + 1^{\text{th}}$ move in δ, and if so, how it is transformed. In other words, the model offers a screenshot of commitments' slates at each point in time but no way to talk about commitments' evolution and the future.

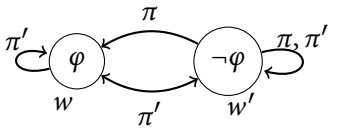
Besides, as the semantics above requires the worlds in π to be transormed by δ **into themselves**, it follows that the propositional nature of the underlying language Φ_0 is essential. Indeed, if we were to adopt a first order dynamic language, instead of Φ_0, a formula in the spirit of $\pi : \neg P(x) \wedge \exists x P(x)$ would wrongfully yield inconsistent commitments (at least in some models).

The only part of states that is 'truly' dynamic is the vector of maximal labels, as it is transformed. The content assigned to labels is not, which has quite unintuitive consequences. For instance, the formula $f = (\pi : \varphi \wedge \neg C_i\psi) \rightarrow (\pi : \psi \rightarrow C_i\bot)$ is a theorem for $\pi \in \Pi_i$: after interpreting $\pi : \varphi$, the semantics ensures that π becomes the label maximal for i, then $\neg C_i\psi$ filters out the states in which the assignment to π verifies (in addition to φ) that ψ. Then, since $\varphi : \psi$ is a test and not a transformation, the only way it can be verified, is that the assignment to π is empty. So i is committed to the absurd. Yet understanding $\pi : \psi$ as a refinement of the content assigned to π rather than a test should make, at least, some states not transformed by f, for some φ and some ψ.

The right thing to do then, is to write a semantics for discourse moves that transforms the assignment of content to labels rather than test them, exactly as assignment of variables to individual of the models are transformed by DRSs or DPL formulae. However, this appears to requires a different setting that the one we have set up in this section. Here is the problem: consider a vocabulary of two labels $\pi \in \Pi_0$ and $\pi' \in \Pi_1$, and the assignment below over a set of two worlds w and w', where the set of worlds assigned to π at world w is symbolized with π-labeled arrows from w to the worlds assigned to π at w. For sake of simplicity, we drop the representation of questions and maximal labels here:

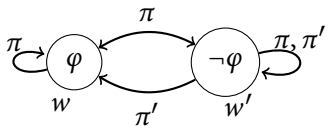


Let us call σ this assignment and place ourselves in the state (w, σ) (again, we drop here the c and $\sim$ for simplicity). What should be the effect of $\pi : (\pi' : \varphi)$ on that state? It should leave the world unchanged and transform σ into an assignment σ'. There are two possibilities. The first is to remove from the assignment to π those worlds in which π' is not assigned only φ worlds. We would thus get:



We see that, in world w, 0 is now committed to an absurdity *via* φ. In the case of a two-labels vocabular, this option makes $\pi : (\pi' : \varphi)$ essentially equivalent to $\pi : C_1\varphi$, but **not** to $C_0C_1\varphi$ (which is what our initial semantics would amounts to). This assymmetry is unsatisfactory: $\pi : \varphi$ is interpreted as an update at the top-level, but as a test when nested in another construction. Why would be the reason for such a behaviour? There is, moreover, another difficulty. We see on the above assignment that, in world w, 1 commits that 0 is committed to 'either $\neg\varphi \wedge C_0\neg\varphi$ or $\varphi \wedge C_0\bot$'. Hence 1 has acknowledged 0's move, but only in one of her commitment-world, because that world 'accidentaly' appeared to be shared with 0. This appeals to a more complex construction.

A second option, to keep with the present notions of worlds, states, assignments and dynamic propositions would be to define $\pi : \delta$ by transforming each of the worlds assigned to π using δ. Let us consider, this time, the update of (w, σ) by $\pi : (\pi' : \neg\varphi)$. We would get the assignment σ' below:



But we see that this has an immediate, extremely unintuitive consequence. Because of the particular structure of the initial assignment σ, using the move $\pi : (\pi' : \neg\varphi)$, 0 is able to change 1's (1st order) commitments and yield an assignment in which $C_1\neg\varphi$, in other word, he can "make 1 talk"!

Both problem stem from the impossibility of our global assignments to dynamically produce a distinction between worlds that are identic in a given initial state. This is a well known problem in dynamic epistemic logic: formalizing private annoucements within a subset of agents expresses the same need for a given action to 'split' worlds. We will see in the next chapter how to apply these techniques to our model.

8.5.3 Conclusions

In this chapter, we have presented new notions of credibility and attacks on credibility that are relevant to the formalization of attacks and credibility threats, in strategic conversations where interlocutor preferences may be opposed. We have developed a symbolic dialogue model extending both Lascarides and Asher (2009) to capture ideas from Stone and Lascarides (2010) and previous sections, with a semantics for dialogue turns and commitments that allows for arbitrary nestings of commitments. We have also shown that this complexity is required to analyze many examples of dialogue with attacks on credibility. We have exposed some limitations of our model that we will now improve on.

Chapter 9

A dynamic logic of ambiguous public commitments

Contents

9.1 Introduction

In previous chapters, we have argued that a semantics in terms of commitments is an attractive idea (a claim also supported by Hamblin, 1987; Traum and Allen, 1994; Traum, 1994). Furthermore we agree with Asher and Fernando (1997); Lascarides and Asher (2009) that dialogues typically involve differing points of view, which the semantics must separately represent. To this picture we have add a complication that has gone so far unnoticed in formal semantics and the prior work we just mentioned, albeit it is well-known from epistemic game theory: commitment slates interact; agents typically commit to the fact that other agents make certain commitments. We have thus formulated a semantics of dialogue moves in terms of nested, public commitments. We have proposed a relational language inheriting the main concepts of SDRT, with a linearized syntax that eases the expression of an SDRS as a succession of discourse moves, keeping everything at a semantic level of analysis where each move's logical form has already been derived (but can involve ambiguities).

This language however, suffers from two limitations: from section 8.5.2 we know that the language lacks a 'true' dynamics, and from the treatment of examples like example (6.3.5) in section 8.3 we know that the language is too permissive, insofar as the consistency of one's commitments about others' commitments is

not tied to the moves that have actually been performed: no assumption has been made of linguistic competence and/or evidence of acknowledgment that would constrain the commitments that one makes about the content of previous moves. Instead we have so far left it up to an abstract entity (the Jury of section 7.4), monitoring the conversation, to decide on the admissibility of a claim about commitments. For instance we have argued in the context of example (6.3.5), that given Quayle's suggestion of being of equal stature as John Kennedy, Bensen's claim that he commited to such a fact was admissible. But we should provide a more systematized account of what one can consistently say and what one cannot, in particular regarding others' moves. Such an account should then be used in the course of the Jury's evaluation.

In this chapter, we will address these two limitations of the model, with two simplifying assumptions: to address the first issue, we will, in a first step, concentrate our efforts on the dynamics of commitments, its logic, and its link to acknowledgments. We will therefore drop the full relational semantics and define a dynamic propositional logic, with only some relational actions (acknowledgments). Recovering the full expressivity of SDRT discourse labels will be the topic of chapter 10. To address the second issue, we will further assume a simplified representation of ambiguity.

Ambiguity arises in dialogue content at various levels of granularity–lexical, syntactic, semantic levels and at the level of discourse structure. In context, these ambiguities trigger pragmatic inferences. These different mechanisms interact in an especially complex way: for instance, the coherence of a dialogue agent i's contribution is tied to the possibility of inferring coherence relations between i's utterances (see chapter 2) which often constrain in return the possible disambiguisations of those utterances, and force or cancel scalar implicatures (Asher, 2013). Integrating every part of such a complex interplay as well as the dynamics of commitments into a single model would be too wide a scope for this chapter. Too keep things manageable, we assume ambiguities of any kind (lexical, syntactic, semantic, or relative to the presence/absence of an implicature as part of a move's content) to be provided externally and, at the semantic level of analysis that we adopt, built into discourse moves of the form $m_1 \sim m_2$, where $\sim$ is an operator of ambiguity, syntactically producing an ambiguous move from a set of moves in the language representing its possible resolutions.

We will attribute to an underlying linguistic theory the task to axiomatise the onset of ambiguity, for instance SDRT's default reasoning and glue logic (section 2.3), relying on axioms in the spirit of $(\pi : \varphi \wedge \pi' : QAP(\pi_o, \pi) \wedge S_i\psi \wedge QAP(\pi_o, \psi)) \rightarrow [(\pi : \varphi) \sim (\pi : (\varphi \wedge \psi)]$[36] (adding $\sim$ to the language of the previous section, but assuming a dynamic interpretation such as chapter 10 will provide); looking back to chapter 8 and the treatment of example (6.3.5), such a mechanism should inform us that Quayle's move suggests that he is of equal stature as Kennedy, and that given the question he was asked, this suggestion makes his commitment ambiguous between a commitment that he has as much experience in the congress as Kennedy did ($\approx$JFK), and another commitment that, in addition to this, he has equal stature that John Kennedy ($\approx$JFK $\wedge$ =JFK), so that *in fine*, his move is represented in the propositional language as $[(\approx \text{JFK}) \sim (\approx \text{JFK} \wedge =\text{JFK})]$. In so doing we also drop the distinction between different strength of commitment, but this could be reconquered (under the same assumptions), with distinct ambiguity constructors $\sim_S$ of different strengths S.

What remains central for the propositional dynamic language to account for is that, by performing a dialogue act X, an agent A commits to some content, potentially ambiguous, then, by responding with another dialogue act Y, a second agent B might commit to having interpreted X in a particular way. (or else to be incoherent). We define a semantics for the propositional dynamic language and provide an axiomatization for a simple, first kind of dynamics. We then consider some problems this simple dynamics has with the semantics of dialogue acts like acknoweldgments, and produce a second dynamics that resolves the problems of the first.

[36] Or, maybe more accurately, an inference rule of the glue logic deriving the ambiguity from clues expressible in the glue language that such a suggestion and question are present.

9.2 A Language for the Dynamics of public Commitment with Ambiguities

9.2.1 Syntax

To model the dynamic of public commitment with ambiguous signals, we assume here an abstract, simplified view of conservations as sequences of $\langle (linguistic)action, speaker\rangle$ pairs. We will build ambiguity into the linguistic actions recursively: in the base case, an action is an unambiguous utterance, whose content we simplify to be a propositional formula. Ambiguous actions are recursively constructed from a set of (lower-level) actions (representing its possible disambiguations). In order to explain this in more formal terms, we introduce some preliminary definitions: Let PROP denote a set of propositional variables (at most countably infinite) and I a set of agents. We define simultaneously the set of actions $\mathcal{A}$ and formulas $\mathcal{L}_o$:

Definition 58 (Actions and formulas)**.** $\mathcal{A}$ and $\mathcal{L}_o$ are the smallest sets such that:

$$\begin{array}{ll}
\forall p \in \text{PROP}\ p \in \mathcal{L}_o & \forall \varphi \in \mathcal{L}_o\ \varphi! \in \mathcal{A} \\
\forall \varphi, \psi \in \mathcal{L}_o\ \forall i \in I\ \neg\varphi, C_i\varphi, \varphi \wedge \psi \in \mathcal{L}_o & \text{for any finite collection of actions } (\alpha_s)_{s=1\ldots n} \text{ in } \mathcal{A} \\
\forall \varphi \in \mathcal{L}_o\ \forall \alpha \in \mathcal{A}\ \forall i \in I\ [\alpha^i]\varphi \in \mathcal{L}_o & (\sim \alpha_s)_{s=1\ldots n} \in \mathcal{A}
\end{array}$$

Additional logical constants and connectors are defined as usual: $\varphi \vee \psi \equiv \neg(\neg\varphi \wedge \neg\psi)$, $\varphi \to \psi \equiv \neg\varphi \vee \psi$, $\bot \equiv p \wedge \neg p$, $\top \equiv p \vee \neg p$.

9.2.2 Semantics

The semantics of our language is based on that for Public Announcements logic (PAL) with private suspicions introduced in Baltag et al. (1998). More specifically, we translate each of our actions in *action structures* of Baltag et al. (1998) and then rely on their semantics.

Recalling some basic definitions, a *frame* is a tuple $\langle W, (R_i)_{i\in X}\rangle$ with W a set of worlds and for each $i \in I$, R_i is a binary relation over W, and a *model* $\mathcal{M}$ is a pair $\langle \mathcal{F}, v\rangle$ with $\mathcal{F}$ a Kripke frame and $v : W \mapsto \wp(\text{PROP})$ an assignment at each world w of propositional variables true at w. We will sometimes use models as superscripts for set of worlds $W^{\mathcal{M}}$, or accessibility relations $R_i^{\mathcal{M}}$ to refer to the set of worlds or the relation of that particular model or frame. We will also abuse notation and write $w \in \mathcal{M}$ as a shortcut for $w \in W^{\mathcal{M}}$. A *pointed model* is a pair $\langle \mathcal{M}, w\rangle$ with $w \in W^{\mathcal{M}}$.

The semantics of action-free formulas is as usual with respect to a pointed model:

Definition 59 (Semantics of static formulas)**.**

$$\begin{array}{l}
\langle \mathcal{M}, w\rangle \vDash p \text{ iff } p \in v^{\mathcal{M}}(w) \\
\langle \mathcal{M}, w\rangle \vDash \neg\varphi \text{ iff } \langle \mathcal{M}, w\rangle \nvDash \varphi \\
\langle \mathcal{M}, w\rangle \vDash \varphi \wedge \psi \text{ iff } \langle \mathcal{M}, w\rangle \vDash \varphi \text{ and } \langle \mathcal{M}, w\rangle \vDash \psi \\
\langle \mathcal{M}, w\rangle \vDash C_i\varphi \text{ iff } \forall w', R_i^{\mathcal{M}}(w, w') \to \langle \mathcal{M}, w'\rangle \vDash \varphi
\end{array}$$

In order to provide a semantics for terms with actions, we need the definition of an *action structure* in Baltag et al. (1998):

Definition 60 (Action Structures)**.** An action structure is a pair $\langle \mathcal{F}, pre\rangle$ where $pre : W^{\mathcal{F}} \mapsto \mathcal{L}_o$ associates to each world in $\mathcal{F}$ a formula, called the *precondition* of this world.

Interpreting formulas with actions require us to first update the model with the action, then to evaluate the formulas with respect to the updated model. As mentioned earlier, we proceed in two steps; we first

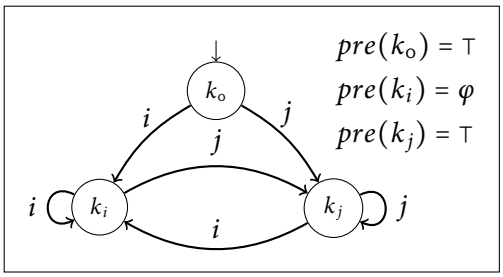
associate a pointed action-structure with each action in $\mathcal{A} \times I$, and then classically update the model with this action. We first recall Baltag et al. (1998)'s informal definition of the update operation. The update of a model $\mathcal{M}$ through an action a is obtained by taking, for each world in a's structure a different copy of $\mathcal{M}$'s world that satisfy the precondition, then allowing a transition for agent i from a world to another iff i)the two worlds were initialy i-related and ii)the two copies they belong to are i-related in a's structure. More precisely:

Definition 61 (Action updates)**.** Let $S = \langle \mathcal{F}, pre \rangle$ be an action structure. Let $k \in S$ and let $\langle \mathcal{M}, w_o \rangle$ be a pointed model. Let $|\varphi|^{\mathcal{M}} = \{w \in \mathcal{M} \mid \mathcal{M}, w \vDash \varphi\}$. If $w_o \notin pre(k)$, the update $\langle \mathcal{M}, w_o \rangle \star \langle S, k \rangle$ *fails*. Otherwise, it is defined as $\langle \mathcal{M}^S, (w_o, k) \rangle$ the model with $W^S = \bigcup_{l \in S} |pre(l)|^{\mathcal{M}} \times l$ as set of worlds, accessibility relations defined as $R_i^{\mathcal{M}^S}((w,l),(w',l'))$ iff i)$R^{\mathcal{M}}(w,w')$ and ii) $R_i^S(l,l')$, and valuations left unchanged *i.e.* $v((w,l)) = v(w)$.

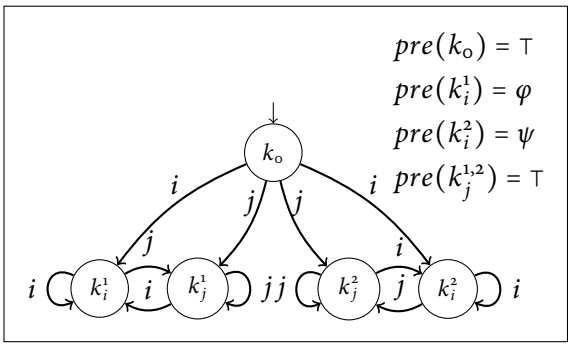
We now provide the translation of conversational moves of our language (*i.e.* elements of $\mathcal{A} \times I$) into pointed action-structures:

Figure 9.1: Some action structures

(a) Unambiguous move $\varphi!^i$



(b) Ambiguous move $(\varphi! \sim \psi!)^i$



A simple unambiguous discourse move by i will generate a *common commitment* to $C_i\varphi$. An ambiguous move, on the other hand, will not but will involve a disjunction of common commitments. A common commitment for a group G towards a proposition φ, $C_G^*\varphi$, has the effect that $C_G\varphi \wedge C_G C_G\varphi \wedge \ldots C_G(C_G)^n\varphi \wedge \ldots$ (analogously to common knowledge). Semantically, we define common commitments for a group G, $C_G^*\varphi$, as

$$\langle \mathcal{M}, w \rangle \vDash C_G^*\varphi \text{ iff } \forall w'(\bigcup_{x \in G} R_x)^+(w,w') \rightarrow \langle \mathcal{M}, w \rangle \vDash \varphi,$$

where the union $\cup$ of two relations is defined as $(R \cup R)(w,w')$ iff $R(w,w')$ or $R'(w,w')$, and R^+ denotes the transitive closure of the binary relation R ($R^+(w,w')$ iff $\exists n > 0 \exists w_1, \ldots w_n \; w_n = w' \wedge R(w,w_1) \wedge \ldots R(w_{k-1},w_k) \ldots R(w_{n-1},w_n)$).

Definition 62 (Interpretation of conversational moves)**.** The interpretation function $[\![\cdot]\!]$ interprets conversational moves of $\mathcal{A} \times I$) as pointed action-structures. Let $m = \alpha!^i \in \mathcal{A} \times I$. $[\![m]\!]$ is defined inductively over α:

- If $\alpha = \varphi!$ then $[\![\alpha]\!] = \langle K, pre, k_o \rangle$ with $K = \{k_o, k_i, k_j\}$, accessibility relation is defined as $R_i^K(k_{\{o,i,j\}}, k_i)$, $R_j^K(k_{\{o,i,j\}}, k_j)$ and no other transitions; preconditions are defined as $pre(k_o) = pre(k_j) = \top$ and $pre(k_i) = \varphi$. The pointed world is k_o and the action-structure is depicted in fig. 9.1(a).

- If $\alpha = (\sim \alpha_s)_{s=1\ldots n}$, let $\langle K^s, pre^s, k_0^s\rangle = [\![\alpha_s!^i]\!]$ be the action structure recursively computed for $\alpha_s!^i$. Assuming the K^s- and $K^{s'}$-worlds are disjoint for $s \neq s'$ (otherwise, first take disjoint copies of the K^ss), define $[\![\alpha_m]\!] = \langle K, pre, k_0\rangle$ with $K = \bigcup_s K^s \setminus \{k_0^s\}$, accessibility relations defined as i)$\forall k \in K^s\, x \in \{i,j\} \quad R_x^K(k_0,k)$ iff $R_x^{K^s}(k_0^s,k)$, ii) $\forall k,k' \in K_s \setminus \{k_0^s\}\, R_x^K(k,k')$ iff $R_x^{K^s}(k,k')$ and iii) there are no other transitions than the one previously listed. pre is defined as $pre(k_0) = \top$ and for $pre(k) = pre^s(k)$ for $k \in K^s$. The pointed world is k_0. fig. 9.1(b) shows the action-structure $[\![(\varphi! \sim \psi!)^i]\!]$ for a move by i which is ambiguous between a commitment to φ and one to ψ.

Note that given this definition, $\sim$ is "associative" in the sense that

$$[\![((\alpha_1 \sim \alpha_2) \sim \alpha_3)^i]\!] = (\alpha_1 \sim (\alpha_2 \sim \alpha_3))^i]\!] = [\![(\alpha_1 \sim \alpha_2 \sim \alpha_3)^i]\!] \text{ (up to renaming of the worlds).}$$

Armed with these definitions, we can now complete the semantics of $\mathcal{L}_0$ providing the semantics for action terms:

Definition 63. Semantics of dynamic formulas:

$$\langle \mathcal{M}, w\rangle \vDash [\alpha^i]\varphi \text{ iff } \langle \mathcal{M}, w\rangle \star [\![\alpha^i]\!] \vDash \varphi$$

Note that due to the fact $[\![\alpha^i]\!]$'s pointed world always has $\top$ as precondition, the update $\langle \mathcal{M}, w\rangle \star [\![\alpha^i]\!]$ cannot fail and the definition is correct.

9.2.3 Worked out example

We illustrate our dynamics by providing an abstract but principled view of the evolving commitments in example (9.2.1):

(9.2.1)a. i : I have my piano lesson in ten minutes. When I get back the shop will be closed.
b. i : And there is no more beer.
c. j : I am not going to get you beer. Go get it yourself.
d. i : I did not say that. I am not asking you to get it.
e. j : Oh yes you did.

What is central to the picture here? i commits to some proposition (we abbreviate it as p), and then to something else, that in its context of utterance might be interpreted as a commitment on a request for j to get beer. i makes an utterance that entails that he takes j to be committed to the request. j then this dispute this commitment of his. i refuses the correction. Assume that we can refer to an external semantic/pragmatic theory that licenses or rejects possible interpretations of a sentence in context, and that such a linguistic theory tells us that (9.2.1-b) as (at least) an assertion that there is no more beer ($\neg b$) licenses a pragmatic inference to a request for i to get some beer ($\neg b \wedge r$). We can then correctly describe example (9.2.1) as involving these action sequences:

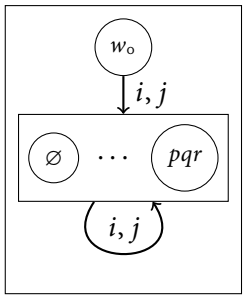
(9.2.2)a. $i : p!^i$
b. $i : (\neg b! \sim (\neg b \wedge r)!)^i$
c. $j : (C_i r)!^j$
d. $i : (C_j C_i r \wedge \neg C_i r \wedge \neg r)!^i$
e. $j : (C_i(C_j C_i r \wedge \neg C_i r) \wedge C_i r)!^i$

Figure 9.2, show how the dynamics transform the initial model. In order to keep the figures readable, we graphically group nodes into clusters, edges going in and out of these clusters are to be understood as distributing over each inner node. We also omit some isolated worlds that therefore have no impact (*i.e.*

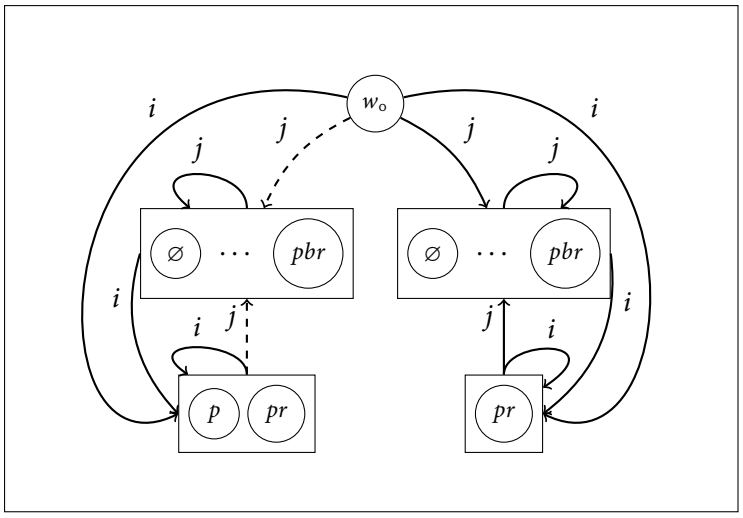
they are present in the definition of action update, but not reachable from any other world). Nodes are labelled by their valuations, except for the actual world labelled as w_0. The initial model of the conversation is depicted in fig. 9.2(a).

Figure 9.2: Models at different stages of example (9.2.2). Arrows should be understood as distributing over all inner nodes.

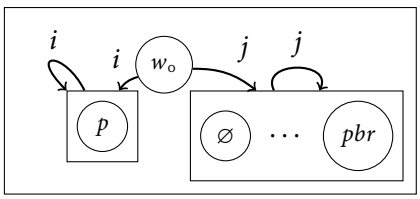
(a) Initial Model for example (9.2.2)



(b) Models after (9.2.2-b) and (9.2.2-c), respectively with and without the dashed edges.



(c) Model after (9.2.2-d)



The initial model in fig. 9.2(a) shows that neither speaker commits to anything. Figure 9.2 shows how i's assertions in (9.2.2-b) to (9.2.2-d) have transformed the commitment space for j and i: after updating with (9.2.2-b), i's public commitments and j's commitments concerning i's commitments are ambiguous as to whether the implicature to go get beer holds; but after the update with (9.2.2-c), only i's commitments remain ambiguous. j's commitments concerning i''s commitments are no longer ambiguous; he commits to i's having committed to the implicature that he should go get the beer. After (9.2.2-d), j's commitments concerning i have become inconsistent. This is a consequence of our strong modeling assumptions about a perfect communication channel leading to common commitments. We will see in section 5 yet another reason to weaken our proposal.

9.3 Complete Deduction System for $\mathcal{L}_0$

One of the interests in keeping the base language of our analysis simple is to be able to investigate the logical properties of the dynamics of commitments. Accordingly, in this section, we present a complete deduction system for $\mathcal{L}_0$. The system and completness proof follow from the general picture drawn in (Baltag et al., 1998), where the authors provide a complete deduction system for the language allowing any kind of action-structure. It turns out however, that the restricted action-structures that are the interpretations of our

conversational moves $\mathcal{A}$(see definition 62) allow nice simplifications, most notably the elimination of any reference to action-structures in the syntactic rules. This allows us to have deduction system for L_o which does not require embedding of L_o into an larger language with additional syntactic constructions. The deduction system is presented on fig. 9.3.

Figure 9.3: Deduction system for $\mathcal{L}_o$. $i, j, x \in I$, $p \in$ PROP

All propositional validities

from $\vdash \varphi \to \psi$ and $\vdash \varphi$ to infer $\vdash \psi$	(MP)
$\vdash [\alpha^i](\varphi \to \psi) \to ([\alpha^i](\varphi) \to [\alpha^i](\psi))$	($[\alpha^i]$-normality)
$\vdash C_i(\varphi \to \psi) \to (C_i(\varphi) \to C_i(\psi))$	(C-normality)
from $\vdash \varphi$ to infer $\vdash C_i\varphi$	(C-necessitation)
from $\vdash \varphi$ to infer $\vdash [\alpha^i]\varphi$	($[\alpha^i]$-necessitation)
$\vdash [\alpha^i]p \leftrightarrow p$	(rw1)
$\vdash [\alpha^i]\neg\psi \leftrightarrow \neg[\alpha^i]\psi$	(rw2)
$\vdash [\alpha^i](\psi_1 \wedge \psi_2) \leftrightarrow [\alpha^i]\psi_1 \wedge [\alpha^i]\varphi_2$	(rw3)
$\vdash [\varphi!^i]C_j\psi \leftrightarrow C_j[\varphi!^i]\psi \qquad$ (for $j \neq i$)	(rw4)
$\vdash [\varphi!^i]C_i\psi \leftrightarrow C_i(\varphi \to [\varphi!^i]\psi)$	(rw5)
$\vdash [\sim(\alpha_s)^i_{s\in S}]C_x\varphi \leftrightarrow \bigwedge_S [\alpha^i_s]C_x\varphi$	(rw6)

In order to proof completeness of the above system, we adapt step by step (Baltag et al., 1998)'s proof to the simplified system. The proof function by reduction of the logic to the static logic K. The idea behind the proof is, once soundness is established, to see our system's axioms as rewrite rules (rewriting the left-hand sides of the equivalences into the right-hand sides), and show that the system is able to proof the equivalence of any given formula to an action-free formula. From there it is quite straightforward to reduce provability of a formula to provability of an action free formula, which is granted as K−axioms are part of our system.

Proposition 2. The deduction rules are sound.

Proof. The proof is trivial for Modus Ponens, necessitation and normality rules. (rw1)'s, (rw2)'s and (rw3) soundness follows directly from definitions (and the fact that our actions never fail). (rw4),(rw5) and (rw6) requires a little more work:

- Let for a world $w \in \mathcal{M}$ $\langle \mathcal{M}^\alpha, (w_o, k_o)\rangle$ denote $\langle \mathcal{M}, w\rangle \star [\![\varphi!^i]\!]$, the update of $\mathcal{M}$ by action φ^i at w (recall that the set of worlds and relations of $\mathcal{M}^\alpha$ does not depend on w). Notice first that the following is true for any formula γ, $k \in \{k_o, k_i, k_j\}$ and w such that $(w, k) \in \mathcal{M}^\alpha$:

  $$\langle \mathcal{M}^\alpha, (w, k_o)\rangle \text{ and } \langle \mathcal{M}^\alpha, (w, k)\rangle \text{ are bisimilar.}$$

  by definition the valuations of (w, k_o) and (w, k) are the same. Since the worlds accessible from k_o in $[\![\varphi]\!]$ are exactly those accessible from k, it follows from the definition of $\mathcal{M}^\alpha$ that the worlds

accessible from (w, k_o) are exactly those accessible from (w, k) which is sufficient to establish the bisimulation.

Let us now prove the soundness of (rw5). The proof for (rw4) is similar. Let $\langle \mathcal{M}, w_o \rangle$ be a pointed model. $\langle \mathcal{M}, w \rangle \vDash [\varphi!^i] C_i \psi$ iff $\langle \mathcal{M}^\alpha, (w_o, k_o) \rangle \vDash C_i \psi$ iff $\forall (w, k) \in \mathcal{M}^\alpha \; R_i((w_o, k_o), (w, k)) \to \langle \mathcal{M}^\alpha, (w, k) \rangle \vDash \psi$. Since we have shown that $\langle \mathcal{M}^\alpha, (w, k) \rangle$ is bisimilar to (w, k_o) and using the definition of $\mathcal{M}^\alpha$, we find the above to be further equivalent to $\langle \mathcal{M}, w \rangle \vDash \varphi$ and $R_i^{\mathcal{M}}(w_o, w) \to \langle \mathcal{M}^\alpha, (w, k_o) \rangle \vDash \psi$. But since by definition $\langle \mathcal{M}, w \rangle \vDash [\varphi!^i]\psi$ iff $\langle \mathcal{M}^\alpha, (w, k_o) \rangle \vDash \psi$, satisfaction of the initial formula is finally equivalent to $\langle \mathcal{M}, w_o \rangle \vDash C_i(\varphi \to [\varphi!^i]\psi)$.

- Let $\alpha = ((\sim \alpha_s)_{s \in S})$ and $x \in I$. by construction, for any world k x-accessible from k_o in $[\![\alpha]\!]$ there is world k_s in $[\![\alpha_s]\!]$ x-accessible from k_o and such that $\langle K^\alpha, k \rangle$ and $\langle K^{\alpha_s}, k_s \rangle$ are bisimilar. This implies that for any world in (w, k) x-accessible from (w_o, k_o) in $\mathcal{M}^\alpha$ there is a $s \in S$ and a world (w, k_s) in $\mathcal{M}^{\alpha_s}$ such that $\langle \mathcal{M}^\alpha, (w, k) \rangle$ and $\langle \mathcal{M}^{\alpha_s}, (w, k_s) \rangle$ are bisimilar. Conversely, for any world $(w, k_s) \in \mathcal{M}^{\alpha_s}$ x-accessible from (w_o, k_o) there is a bisimilar world $(w, k) \in \mathcal{M}^\alpha$ x-accessible from (w_o, k_o).

  Assume that for each α_s, $\langle \mathcal{M}^{\alpha_s}, (w_o, k_o) \rangle \vDash C_x \varphi$. Let (w, k) be x-accessible from $(w_o, k_o) \in \mathcal{M}^\alpha$. We me must have $\langle \mathcal{M}^{\alpha_s}, (w, k_s) \rangle \vDash \varphi$ and by bisimilarity $\langle \mathcal{M}^\alpha, (w, k) \rangle \vDash \varphi$, hence $\langle \mathcal{M}^\alpha, (w_o, k_o) \rangle \vDash C_x \varphi$.

  Conversely, assume that $\langle \mathcal{M}^\alpha, (w_o, k_o) \rangle \vDash C_x \varphi$. Let $(w, k_s) \in \mathcal{M}^{\alpha_s}$ be a world x-accessible from (w_o, k_o), there is a $(w, k) \in \mathcal{M}^\alpha$ bisimilar to (w, k_s) and therefore $\langle \mathcal{M}^{\alpha_s}, (w, k_s) \rangle \vDash \varphi$ and $\langle \mathcal{M}^{\alpha_s}, (w_o, k_o) \rangle \vDash C_x \varphi$.

  All together we can conclude to $\langle \mathcal{M}, w_o \rangle \vDash [(\sim \alpha_s)_{s \in S}] C_x \varphi$ iff $\langle \mathcal{M}, w_o \rangle \vDash \bigwedge_S [\alpha_s] C_x \varphi$, *i.e.* (rw6) is sound.

QED

Proposition 3. Rules (rw1)–(rw6) seen as rewrite rules rewriting the left-hand sides of the equivalences into the right-hand sides form a terminating rewriting system.

This is classicaly obtained from (for instance) the technique of lexicographic path ordering.

Since a rewrite-rule can always be applied to a formula starting with an action, a direct corollary of lemma 3 is that any formula can be rewritten into an action-free formula by the by the rewrite system obtained from the deduction rules.

Proposition 4. If $\vdash \varphi \leftrightarrow \psi$ then for all well formed formula γ of $\mathcal{L}_o$, $\vdash \gamma[\varphi/p] \leftrightarrow \gamma[\psi/p]$.

This can be achieved by induction over the length of γ.

Proposition 13. The deduction system is strongly complete.

Together with lemma 3, lemma 4 yields through a quick induction over the rewrite steps, that for any formula $\varphi \in \mathcal{L}_o$, there is an action-free formula φ_o (one of φ normal forms w.r.t the rewrite system) such that $\vdash \varphi \leftrightarrow \varphi_o$, from there the strong completeness is reduced to the one of modal logic K.

9.4 Acknowledgments and corrections

9.4.1 What is the effect of acknowledgments?

For many researchers, including Clark (1996); Traum (1994); Traum and Allen (1994) *inter alia*, an acknowledgment as in (9.4.1-c) by o of a discourse move m by 1 can signal that o has understood what 1

has said, or that 0 has committed that 1 has committed to a content p with m, and serve to "ground" or to establish a mutual belief that 1 has committed to p. For Clark, grounding by the other conversational participants is a necessary condition for the content of utterances to enter the common ground. Corrections, and self-corrections, as in (9.4.1-d), on the other hand, serve to remove commitments.

(9.4.1)a. 0: Did you have a bank account in this bank?
b. 1: No sir.
c. 0: OK. So you're saying that you did not have a bank account at Credit Suisse?
d. 1: No. sorry, in fact, I had an account there.
e. 0: OK thank you.

The problem is that grounding doesn't follow just from the simple gloss above. With Traum and Allen, we will argue that grounding that i committed to p should require to reach a state where a **common commitment** over some content K holds (*e.g.* typically over the fact that i performed m and the content of m is p). A common commitment by a group of agents G means that every agent in G is committed to K and to the fact that every agent is committed to K, and to the fact that every agent is committed that every agent is committed to K, and so on. But if an acknowledgment performed by agent i only brings additional levels of commitments by i to some given content, it is far from straightforward to see how and why grounding would be possible at all in finite time. Clark indeed mentions such a problem and proposes that the solution lies in a continuous exchange of instantaneous and concurrent, unspoken signals. Other theories such as Lascarides and Asher (2009) assume implicit acknowledgments in the absence of explicit corrections, which might provide another possible explanation of how to fill the gap, for cooperative conversations at least. But there has been no logically precise semantics of acknowledgments that logically entails common commitments and grounding. We provide that below.

Let us get a clearer picture of the problem: We have taken public commitment to be an operator with a weak modal logic (K). We do not believe that commitments validate type 4 or 5 axioms of modal logics; asserting φ does not entail asserting *I assert that* φ; neither does not asserting φ entail asserting *I do not assert that* φ. Indeed Vieu (2011) argues in favor of adding commitments into the semantics of relations in SDRT, insisting that the commitment operator **should not support negative introspection** to account for phenomena such as *e.g.*, contents *blocking* the inference of a coherence relation.

Analogously to common knowledge, let us define common commitments for a group G, $C^*_G\varphi$, as the infinite cunjunction of all possible nesting of the commitment operator for agents in G. $C_G\varphi \wedge C_GC_G\varphi \wedge \ldots C_G(C_G)^n\varphi \wedge \ldots$. If common commitment, even for a single agent, does not follow from static logical axioms, it could follow dynamically, from a background assumption that agents communicate through a perfect communication channel, have a perfect semantic competence implying perfect knowledge and understanding of speakers unambiguous discourse moves and are committed to these two assumptions. This would let the proposition *a message m has been sent* entails the proposition *i commits that a message m has been sent*, for all agents i. From there a common commitment to *m has been sent* would be achieved inductively, via deductive closure of commitments. Hence, the effects of m should be applied at every level on every agent's commitments, and produce a common commitment that the speaker of m commits to the content of m.

Such an assumption underlies the dynamic logic that we developped so far (we will see in section 10.1 how it might be seen to correspond to such an inductive process): it leads to a very strong view of assertions and other discourse moves: our dynamic logic so far has $[\varphi!^i]C^*C_i\varphi$ as a theorem. But then, it appears that grounding acknowledgments are semantically superfluous (we shall formalize this in the dynamic propositional logic, after concluding this general discussion of acknowledgments and common commitments). Moreover, perfect communication channels and perfect linguistic competence are often not realized, even in the most constrained settings. People can argue that they made a sincere, honnest

mistake in their commitment as in (9.4.1-d), or they can (sincerely or not) act as if they were unaware of such a mistake, in denying some previously made commitment.

Then, the only informative contribution of an acknowledgment by i of a move by j would be that i agrees with the content of j's move, in other words acknowledgments would always be accept moves, we think this is innacurate: we can imagine 1 in example (9.4.1) acknowledging 0's response even if he patently does not believe its content and neither commits to that answer being true. Such acknowledgments are often present in legal questioning but in many other conversations too.

We might thus opt for a much weaker semantics for discourse moves, where, in particular, an assertion of φ by speaker i only entails that $C_i\varphi$. A similarly weak semantics for acknowledgments would mean that an acknowledgment by j of the content of an assertion by i entails $C_jC_i\varphi$. But this makes grounding impossible in finite conversations: if a discourse move m by i entails only C_ip, (a) and, (b) entails that all the conversational participants believe C_ip (Traum and Allen, 1994; Ginzburg, 2012). Then j's acknowledgment of m would entail $C_jC_ip \wedge Bel_GC_jC_ip$. and Bel_GC. Thinking about conversations as infinit strings of discourse moves in $\mathcal{A}^\omega$ by players i and j, we have, inductively:

Observation 3. For a conversational sequence $\sigma \in \mathcal{A}^\omega$ of assertions and acknowledgments, no finite prefix $\sigma_k \in \mathcal{A}^k$ of σ of arbitrary length k ever entails a common commitment to the fact that i commits to φ, $C^*C_i\varphi$.

That is, common commitments might be achieved only after an infinite sequence of acknowledgment moves between i and j.

Some other options are logically possible. For instance, we could keep the simple semantics for discourse moves, but assign acknowledgments a very strong semantics. On such a semantics, j's acknowledging a discourse move by i that entails $C_i\varphi$ would imply a common commitment by i and j to $C_i\varphi$. But this implausibly imputes to j the ability to force commitments on i that quickly leads to absurdities. If i says, for instance, *I don't want to go to the meeting*, j can "acknowledge" i's move by saying, *OK, thank you very much for agreeing to go to the meeting*, thus forcing a common commitment to C_i *i goes to meeting*. But clearly i didn't commit to going to the meeting.

9.4.2 Do agents need to achieve common commitments?

Can we do without common commitments? We think not. Common commitments are essential (see also Clark (1996)) for strategic reasons and can be present even when mutual beliefs about a shared task are not.

Suppose, for instance, that i want to adopt a 'fix point' goal in which

1. $C_j\varphi$ and

2. j must 'admit' her defeat, *i.e.* j cannot consistently assert that i fails to achieve her goal.

If i adopts the simplest goal of satisfying the first condition, and extracts from j a move m implying $C_j\varphi$, j has a winning strategy for denying i's victory. She simply denies committing to φ (*I never said that*), since $C_j\neg C_j\varphi$ is consistent with $C_j\varphi$, even if $Bel_jC_j\varphi$. Player j lies, but she is consistent. If i now adopt a more complex goal to achieve $C_j\varphi \wedge C_jC_j\varphi$, j can still similarly counter i's goal while maintaining consistency: j can assert something to the effect that $C_j\neg C_jC_j\varphi$. However, if i adopts achieves a goal of reaching *common commitment* $C^*_GC_j\varphi$, with G the group of conversational participants, j does not have a way of denying her commitment without becoming inconsistent, as $C^*C_j\varphi \rightarrow (C_jC_j\varphi \wedge C_jC_jC_j\varphi \wedge \cdots)$, for any finite depth of nesting of C_j operators. And only common commitments rule out other, more elaborate ways of defeating conversational goals. For instance, on a weaker semantics j could deny that i had committed to what j

had committed to at some level of embedding. Thus, if i's goal depends on j's committing to one of i's commitments, a lack of common commitment will allow j to deny i's achieving her conversational goals.

This argument might seem abstract, but we think that the need for grounding is quite real. In section 6.4 we have made a case that there is an important difference between being non-Gricean and admitting to being so. Assume a view of Gricean SINCERITY in which $Bel_i C_i \varphi \rightarrow Bel_i \varphi$ and a view of LINGUISTIC COMPETENCE in which $C_i \varphi \rightarrow Bel_i C_i \varphi$. Let us revisit the previous argument and assume that j is commited to $C_j^k \varphi$ for $1 \leq k \leq n$ (where C_j^k denotes the nesting of k C_j operator), but not to no higher nesting of commitments. j can not only commits to $\neg C_j^k \varphi$ *consistently*, she can do it and respect SINCERITY as well, at a (rather cheap) cost of committing to have failed once at LINGUISTIC COMPETENCE. Moreover, assuming that each level k of commitment over φ was achieved using a different dialog move, it it is possible that j honnestly failed k times at LINGUISTIC COMPETENCE, then it is possible that j believes that she did not commit to φ. Now if there is a common commitment over $C_j \varphi$, a sincere denial of any level of commitment implies for j to fail LINGUISTIC COMPETENCE for infitely many different propositions which seem intuitively much worse. Of course, the higher k grows, the more difficult it seems to find a real example where one can argue that he failed to understand something k times in a row. But then, using a jump to common commitment is a way, in a symbolic setting, to model a point where there is no more plausibly sincere rebutal.

We could have formulated the problem of grounding and common commitments in doxastic terms, *i.e.* asking that j's acknowledgement of a move by i leads to a mutual belief by i and j in some content. However problems for the semantics of assertions and acknowledgments analoguous to those we have just sketched in terms of commitments would surface for an account of grounding in terms of mutual belief as well. The simple reason for that is that, even in cooperative settings, beliefs changes must themselves reflect some changes in commitments. In cooperative settings though, Gricean cooperativity might come in as an additional piece of machinery allowing to derive inifitely many implicit acknowledgments: Clark and Brennan (1991) observes that grounding seem to require conversationalist to give infinitely many positive bits of evidence–*Requiring positive evidence of understanding seems to lead to an infinite regress*, and claims that some form of evidence such as *continued attention* solves the situation as it can occur continuously and does not require a separate presentation. Lascarides and Asher (2009) assumes implicit acknowldegments in the absence of explicit correction. In both cases, a solution (relying on some form of cooperativity) is proposed where infinitely many acknowledging moves might happen in a finite amount of time. We will propose a formal analysis compatible with these two solutions, with an additional mechanism to allow grounding in non-cooperative settings under certain assumptions.

Let us finally discuss briefly the semantics of self-corrections. While Lascarides and Asher (2009) give a general semantics for corrections in terms of simple commitments, according to which one speaker commits to the negation of what another speaker committed to with a prior move, they do not look at self-corrections: in self-corrections, speakers can not only deny prior commitments but also "undo" or "erase" them. For instance, if in (9.4.1-b) 1 commits to not having a bank account; in (9.4.1-d) 1 no longer has this commitment. Self-corrections thus entail a revision of commitments. No one has proposed a logical analysis of self-corrections. We will very briefly sketch such an analysis and argue that these moves have an essential, strategic role to play in dialogue, even if we assume a perfect communication channel and unambiguous commitments in dialogue moves.

9.4.3 Acknowledgments in the dynamic commitment logic

To treat acknowledgments, we first enrich our language into a language $\mathcal{L}_D$ with actions for acknowledgments. We do that by adding the recursive construction $Ack(\alpha^x)$ to the set of linguistic action $\mathcal{A}$, for any $\alpha \in \mathcal{A}$ and $x \in I$. Defining the semantics of $\mathcal{L}_D$ only requires us to define the interpretation of acknowledgment-actions into action-structures. Let $\alpha \in \mathcal{A}$ be a linguistic action. Let $\langle K^\alpha, k_o^\alpha, pre^\alpha \rangle = [\![\alpha^x]\!]$.

Let k_o and k_j be "fresh" symbols not appearing in K^α.

Definition 64 (Acknowledgment as action structures)**.**

$$[\![Ack(\alpha^x)^i]\!] = \langle \{k_o, k_j\} \cup K_\alpha, k_o, pre \rangle$$

Accessibility relations are defined as $R_i(k_o, k_o^\alpha)$, $R_j(k_o, k_j)$, $R_{i,j}(k_j, k_j)$, $\forall k, k' \in K^\alpha, \forall x \in \{i, j\}$ $R_x(k, k')$ iff $R_x^{K^\alpha}(k, k')$ and no other transitions. $pre(k_o) = pre(k_j) = \top$ and pre coincide with pre^α on K^α.

Acknowldegments are relational moves (actually, the only relational moves of our dynamic language). This means that conversational agents should not have the action $Ack(\alpha)$ for just any α as part of their vocabulary, but might make such a move only if α is a move which is part of the conversational context, *i.e.* has been played by an agent as such, or as part of an ambiguous action. We do not focus on imposing such constraints into the language itself, but a fully relational account of acknowledgments is proposed in chapter 10.

It is easy to check that effects of action $Ack(\alpha^x)^i$ commit i to the effects of α^x and that, given the dynamics of sections section 9.2 and section 9.3, acknowledgments of previous actions have no effect in the sense that $\langle \mathcal{M}, w \rangle \vDash [\alpha^x][Ack(\alpha^x)^i]\varphi$ iff $\langle \mathcal{M}, w \rangle \vDash [\alpha^x]\varphi$. This formalizes the problem. To address the problem, we provide an alternative *weak* semantics for $\mathcal{L}_D$, in which we redefine the interpretation of linguistic actions as action structures. Only unambiguous utterance-actions need a new definition, as the recursive computation mechanism of action-structures for ambiguous utterances- and acknoledgments-actions stays the same.

Definition 65 (Weak action interpretation)**.** Define $[\![\cdot]\!]^w$ by $[\![\varphi!^i]\!]^w = \langle k_o, k_i, k_1, k_o, pre \rangle$ with $R_i(k_o, k_i)$, $R_j(k_o, k_1)$, $R_{i,j}(\{k_i, k_1\}, k_1)$ and no other transitions. $pre(k_o) = pre(k_1) = \top$ and $pre(k_i) = \varphi$

$$[\![\sim(\alpha_s)_{s\in S}^i]\!]^w \text{ and } [\![Ack(\alpha^x)]\!]^w \text{ are computed as before}$$

Interpretation of assertions under the strong and weak semantics, and interpretation of acknowledgments thereof are depicted on fig. 9.4

Define finally $\vDash^w$ as the new truth-maker operator defined as $\vDash$ was, but this time based on the interpretation $[\![\cdot]\!]^w$ of linguistic actions.

Under $\vDash^w$ action $[\varphi!^i]$ has i commits to φ, but changes neither i's second order commitments (in general $\langle \mathcal{M}, w \rangle \nvDash^w C_i C_i \varphi$) nor anyone else's commitments. This now fixes our problem of the liar who denies commitments he has previously made; someone can now commit to φ but then later say *I never said* φ and remain consistent.

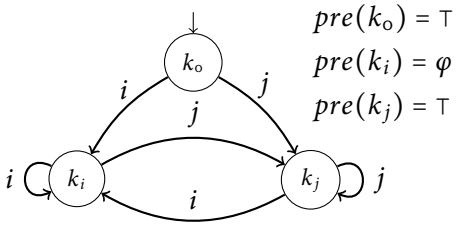
This weaker semantics, however, makes grounding impossible in finite conversations. We think nevertheless that grounding is possible. Our proposal is that a particular sort of acknowledgment and confirming question licenses the move to common commitment. It is the one in (9.4.1-b) and (9.4.1-c), where o asks a confirming question after an acknowledgment of a move m. If 1's answer to the confirming question is consonant with m, then $C^*_{\{o,1\}} C_1 \varphi$, and o has achieved her goal.

We can formalize this proposal using our notion of ambiguous commitments. An acknowledgment is in fact ambiguous. One reading comes from our simple semantics where an acknowledgment adds one layer of commitment–i.e. if j acknowledges i's commitment to φ with a simple *OK*, we have $C_j C_i \varphi$. The other reading is that it indeed implies a common commitment of the form $C^*_{i,j} C_j \varphi$, following our second semantics for assertions. The clarification question, when answered in the affirmative, selects the common commitment formulation. Our proposal is compatible but distinct from Clark's (ours is also formally worked out), and interestingly survives in non-cooperative settings.

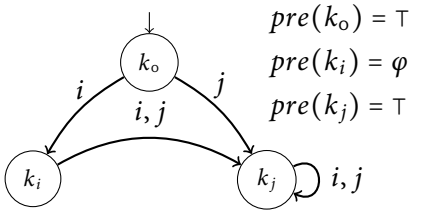
We quickly now turn to corrections. Speakers can not only deny prior commitments but also "undo" or "erase" them with *self-corrections*. For instance, if in (9.4.1-b) 1 commits to not having a bank account; in

Figure 9.4: Assertions under the strong and weak semantics and acknowledgments thereof.

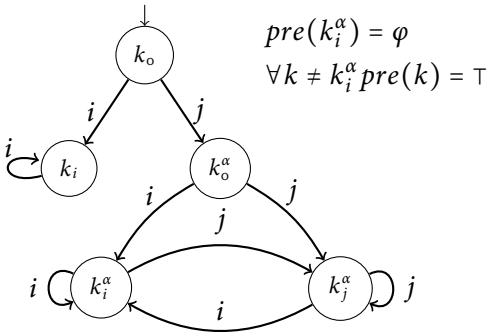
(a) Assertion of φ by i under the strong semantics



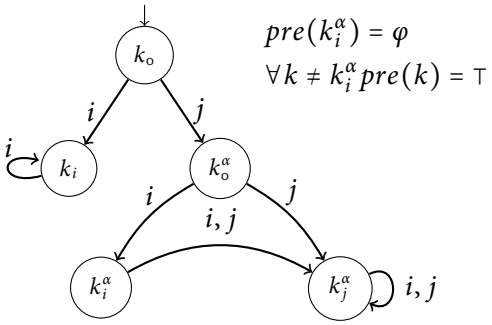
(b) Assertion of φ by i under the weak semantics



(c) Acknowledgment of the (strong semantic) assertion of $\varphi!^i$



(d) Acknowledgment of the (weak semantic) assertion of $\varphi!^i$



9.4.1-d 1 no longer has this commitment (See Ginzburg 2012 for a detailed account of repair). Conversational goals of the form $C_G^* C_i p$ are unstable if i may correct herself; they may be satisfied on one finite sequence but not by all its continuations. j's being able to correct a previous turn's commitments increases the complexity of i's goals (Serre, 2004), which affects the existence of a winning strategy for i; an unbounded number of correction moves will make any stable $C_G^* C_i p$ goal unattainable, if p is not a tautology. We observe, however, a sequence of self-corrections is only a good strategy for achieving j's conversational goals if she is prepared to provide an explanation for her shift in commitments (and such explanations must come to an end). As chapters 7 and 8 argue, conversationalists are constrained to be credible in a certain sense if they are to achieve their conversational goals. Constantly shifting one's commitments with self-corrections leads to non-credibility, thus avoiding the problem of unbounded erasures.

To provide a semantics for corrections, we begin from Lascarides and Asher (2009), who provide a *syntactic* notion of revision over the logical form of the discourse structure. Using the correction of m as an action update on the commitment slate prior to m yields a semantics for corrections. Our formal semantics captures the dynamic effects of announcements, corrections and acknowledgments; common commitments are important conversational goals and that particular conditions must obtain if they are to be achieved.

9.5 Conclusions

We have presented two semantics for dialogue in terms of commitments that is general enough to handle non-cooperative and cooperative dialogues. The first one is conceptually simple and has a straightforward axiomatization but fails to give a sensible semantics for acknowledgments and is also too restrictive concerning denials of commitments, which our semantics makes inconsistent instead of simply a lie. Finally,

we discussed corrections as another problem for the semantics of dialogue and offered a solution.

Chapter 10

Commitments, acknowledgments and grounding in SDRT with nested commitments

Contents

10.1 Introduction

The previous chapter has introduced a dynamic propositional language with ambiguity and acknowledgments. This language is equipped with two distinct interpretations: one, the *strong* semantics, can arguably be thought of as implementing a vision of perfect linguistic competence, communication channel and commitment thereon, and produces as a common commitment over $C^*C_i\varphi$ as the result of an assertion $\varphi!^i$ of φ by i. As a consequence, acknowledgmens are also semantically superfluous under the strong interpretation. The other, the *weak* semantics, let $\varphi!^i$ alter only i's first order commitments, leaving everything else unchanged.

The concern of this chapter is twofold:

- First, we will show how the weak and strong semantics relate, *via* infinite sequences of acknowledgments, confirming our intuition that the strong semantics implements a view of inductive inference of acknowledging moves.
- Second, we will complete the solving of chapter 8's puzzle, turning our dynamic propositional logic into a fully dynamic alternative to the language of chapter 8, with relational predicates ranging over discourse moves.

10.2 Linking the strong and weak semantics for assertions

Going beyond chapter 9, we now provide a link between the strong and weak semantics, on which the strong semantics corresponds to an idealization of an infinite sequence of acknowledgments. In what follows, C^* is a shortcut for C^*_I *i.e.* common commitment over all agents. We first define such sequences:

Definition 66 (Iterated acknowledgments)**.** Let $\alpha^x \in A \times I$ be an ⟨action, agent⟩ pair, let $G \subseteq I$ be a group of agents. Define inductively the set of ⟨action, agent⟩ pairs $\uparrow^n_G (\alpha^x)$ as follows:

- $\uparrow^0_G (\alpha^x) = \{\alpha^x\}$
- $\uparrow^{n+1}_G (\alpha^x) = \bigcup_{i \in G} \bigcup_{\beta \in \uparrow^n_G} ack(\beta)^i$.

Level n in the above hierarchy consists of every possible acknowledgment by every agent in G of actions at the lower level. Note that $|\uparrow^n_G (\alpha^x)| = |G|^n$. An easy induction shows that at each level n, the order in which actions in $\uparrow^n_G (\alpha^x)$ are applied to a model $\mathcal{M}$ do not change the final result; for any complete orderings $\beta_0, \beta_1 \dots \beta_{|G|^n}$ and $\beta'_0, \beta'_1, \dots \beta'_{|G|^n}$ of $\uparrow^n_G (\alpha^x)$, and for every model $\mathcal{M}$, $(\dots((\langle \mathcal{M}, w\rangle \star \beta_0) \star \beta_1) \star \dots) \star \beta_{|G|^n})$ and $(\dots((\langle \mathcal{M}, w\rangle \star \beta'_0) \star \beta'_1) \star \dots) \star \beta'_{|G|^n})$ are isomorphic. Our choice of ordering of $\uparrow^n_G (\alpha^x)$ will thus not affect our results.

Let $<_I$ be any total order on the set of agents I. We inductively extend $<_I$ to $\uparrow^{n+1}_G (\alpha^x)$, by defining $ack(\beta)^i <_I ack(\beta')^j$ iff $i <_I j$ or $i = j$ and $\beta <_I \beta'$ (thus taking a lexicographic ordering at each step). We write $\uparrow^{n+1}_G (\alpha^x)_k$ as the k^{th} action in $\uparrow^{n+1}_G (\alpha^x)$ according to $<_I$. We can now define the infinite sequence of actions

$$!^\omega_G(\alpha^x) = \langle \uparrow^0_G (\alpha^x)_1, \uparrow^1_G (\alpha^x)_1, \uparrow^1_G (\alpha^x)_2, \dots, \uparrow^n_G (\alpha^x)_1, \dots, \uparrow^n_G (\alpha^x)_{|G|^n}, \dots \rangle$$

and

$$ack^\omega_G(\alpha^x) = \langle !^\omega_G(\alpha^x)_n \rangle_{1 \leq n < \omega}.$$

We have defined here two infinite sequences of actions: $!^\omega_G(\alpha^x)$ corresponds to the sequence starting with action α executed by x followed by its acknowledgment by every agent in G, again followed by acknowledgments of these acknowledgments by every agent in G and so on, while $ack^\omega_G(\alpha^x)$ corresponds to the same sequence of acknowledgments without the initial performance of α by x.

The last ingredient needed to complete the correspondence between the strong and weak semantics of acknowledgments is a semantics for such infinite sequences:

Assume in what follows that for any $G \subseteq I$, C^*_G is **not** part of the language $\mathcal{L}_D$[37].

Definition 67. Let $\mathfrak{U}$ be a pointed model and $\langle \mathfrak{U}_n \rangle_{n \in \omega}$ be an infinite sequence of models. $\mathfrak{U}$ is a *limit model* of $\langle \mathfrak{U}_n \rangle_{n \in \omega}$ iff $\forall k \in \omega \exists n_0 \in \omega \forall n \geq n_0$ $\mathfrak{U}$ is k-bisimilar[38] to $\mathfrak{U}_n$. Notice that existence of a limit model requires that $\forall k \exists n_0 \forall n > n_0$ $\mathfrak{U}_{n+1}$ is k-bisimilar to $\mathfrak{U}_n$. Let $\equiv_\omega$ denote unbounded finite bisimilarity, *i.e.* k-bisimilarity for every $k \in \omega$. By transitivity of finite bisimulation, two limit models are $\equiv_\omega$ equivalent, and conversely an $\equiv_\omega$ equivalent to a limit model is a limit model.

Definition 68. Let $\langle \alpha^{x_k}_k \rangle_{k \in \omega}$ be an infinite sequence of ⟨action, agent⟩ pairs. Let $\varphi \in \mathcal{L}_D \cup \{C^*_G \varphi \mid G \subseteq I\}$, and $\langle \mathcal{M}, w\rangle$ be a pointed model. Define inductively $\mathfrak{U}_0 = \langle \mathcal{M}, w\rangle$ and $\mathfrak{U}_{k+1} = (\mathfrak{U}_k \star [\![\alpha_k]\!]^{x_k})$. The effects of an infinite sequence of actions are defined as follows:

- If $\langle \mathfrak{U}_k \rangle_{k \in \omega}$ has no limit model, the effects are undefined.
- Otherwise $\langle \mathcal{M}, w\rangle \models [\langle \alpha^{x_k}_k \rangle_{k \in \omega}]\varphi$ iff $\forall \mathfrak{U}$ if $\mathfrak{U}$ is a limit model of $\langle \mathfrak{U}_k \rangle_{k \in \omega}$ then $\mathfrak{U} \models \varphi$

The following proposition links the strong and the weak semantics:

Proposition 14.

- For an action α, the effects of $ack^\omega(\alpha)$ are always defined and
- let $\langle \mathcal{M}, w\rangle$ be a pointed model, $\langle \mathcal{M}, w\rangle \models [\varphi!^x]\psi$ in the strong semantics for utterances iff $\langle \mathcal{M}, w\rangle \models [!^\omega_I(\varphi!^x)]\psi$ in the weak semantics.

[37] which does not prevent us to study whether actions might yield a model where it holds

[38] For an introductionto finite bisimulation see for instance Blackburn et al. (2006)

We now have a formally precise statement of the intuition that the strong interpretation of $\varphi!^x$ is logically equivalent to $[\varphi!^x][ack_I^\omega(\varphi!^x)]$, *i.e.* that the strong semantics logically behaves as if utterances were always followed by every possible acknowledgment. It also shows something non-trivial: such infinite iterations of acknowledgments are regular enough to be finitely representable as an update of the model. This means that we can nest such constructions without further effort. Following the same kind of reasoning we can then define the action $ack^i(!_I^\omega(\varphi!^x))$, to represent i's acknowledgement of an infinite sequence of acknowledgment yielding a common commitment over $C_x\varphi$, and shows that this action does the same as the "finite" one $ack^i(\varphi!^x)$ under the strong semantics. Our analysis regarding infinite constructions supports our formal treatment of grounding.

We have requested C^* to not be a part of the agents' vocabulary. This might seem an annoying limitation preventing us to use the jump to common commitment of chapter 9 with infinite acknowledgments. Recall that the proposed solution to reach common commitment is to have an ambiguous action between a 'regular' acknowledgment and a 'strong' acknowledgment followed by a clarification question establishing whether or not the common commitment has been reached.

Using infinite actions, the ambiguous acknowledgment of $\varphi!^x$ becomes $[ack^i(\varphi!^x)] \sim [ack_I^\omega(\varphi!^x)]$, which is fine, however a response to the clarification question $C^*\varphi?$ should yield either an action $(C^*\varphi)!^x$ or $(\neg C^*\varphi)!^x$, which we have ruled out. There are different possible ways to address this:

- The initial motivation to rule C^* out of the vocabulary is that an infinite sequence of actions that does not fail on a model $\mathcal{M}$ defines a unique update of the model up to $\equiv_\omega$. It is well known that satisfaction of basic modal formulae is invariant up to $\equiv_\omega$, but that a formula like $\neg C_i\neg C_i^*\neg C_i\bot$ is not because it imposes at least an infinite branch, and we can find two $\equiv_\omega$ inifinitely branching models, one with an infinite branch and one without. However the definitions remain valid, the effect of adding C^* to the vocabulary is to introduce non-determinism: for all we know, there can be an infinite sequence of actions $\langle\alpha_k^{x_k}\rangle_{k\in\omega}$ yielding at least two limit models not equivalent up to full bisimulation, so that $\langle\mathcal{M},w\rangle \not\models [\langle\alpha_k^{x_k}\rangle_{k\in\omega}]\neg C_x\neg C_x^*\neg C_x\bot$ and $\langle\mathcal{M},w\rangle \not\models [\langle\alpha_k^{x_k}\rangle_{k\in\omega}]\neg C_x\neg C_x^*\neg C_x\bot$. But this not per se a problem to our account, in particular we still have:

  $$\models [ack^i(\varphi!^x) \sim ack_I^\omega(\varphi!^x)][(C^*C_x\varphi)!^x][ack((C^*C_x\varphi)!^x)^y]C^*C_x\varphi.$$

  Hence the restriction can simply be raised without much harm.

- Alternatively, since we are only interested in infinite sequences of acknowledgments, we can notice that the sequence of updates of a model through actions in $ack_I^w(\alpha)$ exhibit more structure than we have exploited so far. In fact, we have exploited in our definition that iterated acknowledgments successively refine commitments of higher order, and therefore yield sequences of model such that the $k+1$th is k-bisimilar to the k-th. But there is more, actions as we have considered them cannot 'create' an infinite path out of nothing. They can unfold a path, duplicate it, and eventually prune it. So we might update the definition of limit models, for instance, in the following way: $\mathfrak{U}$ is a limit model iff of $\langle\mathfrak{U}_n\rangle$ iff there is a sequence of simulations G_k of $\mathfrak{U}_{k+1}$ by $\mathfrak{U}_k$ such that G_k is also a k-bisimulation between the two pointed model, and in addition there is a sequence F_k of simulations of $\mathfrak{U}$ by the $\mathfrak{U}_k$ such that $F_{k-1} = F_k \circ G_k$. With this more demanding notion of a limit model, we can show that there is at most a unique limit model up to (full) bismilarity.

As already mentioned, existing theories, without proposing a fully worked out semantics, solve the problem of grounding by assuming that, at least by default, infinitely many weak acknowledgments can synchronously be performed by conversational agents. The previous discussion establishes that the strong semantics implements precisely such a behavior. Our proposed solution thus complements existing ones.

10.3 Semantics with nested commitments for richer languages

In the preceding sections, we gave a semantics for acknowledgments using a propositional language in which discourse moves like acknowledgments were analyzed as action operators on propositions. While this approach works for acknowledgments, it is difficult to adapt this to truly relational moves like `Correction` or most discourse relations. In this section we offer a full dynamic semantics for a language with nested commitments in which we can refer to conversational moves and make assertions about relations between conversational moves.

We build on the mechanisms of SDRT (Asher and Lascarides, 2003), with some modifications and extensions: we deal with nested commitments by allowing the informative content of discourse labels to talk about the content assigned to other discourse labels. As in SDRT, we will assume given lower-level language, typically, Dynamic Predicate Logic (DPL, Groenendijk and Stokhof (1991)), or DRT (Kamp and Reyle (1993)), and extend this language with discourse labels and relations.

Let I be a finite set of agents. Let $\Pi = \biguplus_{i\in I} \Pi_i$ be a disjoint union of $|I|$ countable sets of symbols (discourse labels). By definition, agent i is the *speaker* of any label $\pi \in \Pi_i$, which we write $spk(\pi) = i$. Let $\mathcal{R}$ be a finite set of relation symbols.

We assume a basic language $\mathcal{L}_o$ with the following: $\mathcal{L}_o$ has a binary constructor $\wedge$ and a unary constructor $\neg$, *i.e.* $\forall \varphi, \varphi' \in \mathcal{L}_o\ \varphi \wedge \varphi', \neg\varphi \in \mathcal{L}_o$. We assume a class of models for $\mathcal{L}_o$, and every model $\mathcal{M}$ defines a set of states $X^{\mathcal{M}}$ and an interpretation $[\![\cdot]\!]_o^{\mathcal{M}}$ such that $[\![\varphi]\!]_o^{\mathcal{M}} \subseteq X \times X$, $[\![\neg\varphi]\!]_o^{\mathcal{M}} = \{(x,x) \mid \neg\exists x'(x,x') \in [\![\varphi]\!]_o^{\mathcal{M}}\}$ and $[\![\varphi \wedge \psi]\!]_o^{\mathcal{M}} = [\![\varphi]\!]_o^{\mathcal{M}} \circ [\![\psi]\!]_o^{\mathcal{M}} = \{(x,y) \mid \exists z\,(x,z) \in [\![\varphi]\!]_o^{\mathcal{M}} \text{ and } (z,y) \in [\![\psi]\!]_o^{\mathcal{M}}\}$. The latter requirements ensure that the interpretation of $\neg$ and $\wedge$ coincide at the lower and upper levels.

Our intended base-level language is DPL, as it yields a rich and expressive discourse semantics that can handle questions. However, in order to make clear the link between the present section's semantics and the dynamic logic of the previous section, we will also occasionally use a simpler propositional test logic $\mathcal{L}_{\texttt{test}}$ (which can be seen as a fragment of DPL). Simply define $\mathcal{L}_{\texttt{test}}$ as to contain the same formulae as the propositional logic over signature PROP, states $X_{\texttt{test}}$ are valuations PROP $\mapsto \{0,1\}$, and for two valuations $v, v' \in X_{\texttt{test}}$, define $[\![\varphi]\!]_{\texttt{test}} = \{\langle v, v\rangle \mid v \vDash \varphi$, where $\vDash$ is the classical truth-maker of the (static) propositional logic over PROP.

The language $\mathcal{L}_{\Pi,\mathcal{R}}$ extends $\mathcal{L}_o$ and is defined by structural induction:

Definition 69. $\mathcal{L}_{\Pi,\mathcal{R}}$ is the smallest language such that:

$$\begin{aligned}
&\forall \varphi \in \mathcal{L}_o\ \varphi \in \mathcal{L}_{\Pi,\mathcal{R}}\\
&\forall \pi_1, \pi_2, \pi_3 \in \Pi\ \forall R \in \mathcal{R}\ R(\pi_1,\pi_2,\pi_3) \in \mathcal{L}_{\Pi,\mathcal{R}}\\
&\forall \alpha \in \mathcal{L}_\pi\ \forall i \in I\ C_i\alpha \in \mathcal{L}_{\Pi,\mathcal{R}}\\
&\forall \alpha \in \mathcal{L}_\pi\ \forall \pi \in \Pi\ \pi : \alpha \in \mathcal{L}_{\Pi,\mathcal{R}}\\
&\forall \alpha, \alpha' \in \mathcal{L}_\pi\ \alpha \wedge \alpha' \in \mathcal{L}_{\Pi,\mathcal{R}}\\
&\forall \pi, \pi_1, \pi_2 \in \Pi\ \pi_1 \overset{\pi}{\sim} \pi_2 \in \mathcal{L}_{\Pi,\mathcal{R}}\\
&\forall \alpha \in \mathcal{L}_\pi\ \neg\alpha \in \mathcal{L}_{\Pi,\mathcal{R}}
\end{aligned}$$

An SDRS is a formula of $\mathcal{L}_{\Pi,\mathcal{R}}$ such that every occurrence of a formula of $\mathcal{L}_o$, or a formula of the form $C_i\alpha$ is guarded by some label $\pi \in \Pi$. In addition, define $\varphi \to \psi$ as $\neg(\varphi \wedge \neg\psi)$.

$\mathcal{L}_{\Pi,\mathcal{R}}$ adds several constructions to $\mathcal{L}_o$. First there are statements of the form $\pi : \varphi$ that "store" proposition φ in the content of a discourse label π. It implements the linguistic action of the speaker of π's saying φ. φ may itself be a structured discursive object, as in SDRT. Relational propositions $R(\pi_1,\pi_2,\pi_3)$ update the content of π_3 to express a relational move, recursively computed through the meaning of R and the content assigned to π_1 and π_2. Commitment operators are added, as well as a construction $\pi_1 \overset{\pi}{\sim} \pi_2$, which will

implement ambiguity between the discourse moves stored under π_1 and π_2. We now turn to the semantics of $\mathcal{L}_{\Pi,\mathcal{R}}$.

A model for $\mathcal{L}_{\Pi,\mathcal{R}}$ is simply a model $\mathcal{M}$ for $\mathcal{L}_0$. Let $C = \{c \in (\Pi)^I \mid \forall i \in I\ c(i) \in \Pi_i\}$, the set of function associating to each agent i a label $c(i)$ whose speaker is i. The set of ($\mathcal{L}_{\Pi,\mathcal{R}}$-)states $S^{\mathcal{M}}$ for $\mathcal{M}$ is the product $X^{\mathcal{M}} \times C \times F^{\mathcal{M}}$ where $F^{\mathcal{M}}$ is the set of assignments to labels (also called label-assignments in the remainder of the chapter). A state for the extended language $\mathcal{L}_{\Pi,\mathcal{R}}$ is thus a triple, consisting of a state x of the base language, a label $c(i)$ for each agent representing her current commitments[39], and a label-assignment f. We define label-assignments below. The purpose of a label-assignment f is to assign a (semantic) proposition $f(\pi)$ to a given discourse label π—*i.e.*, in our dynamic logic, a set of transitions between states. Such a transition is technically a pair $\langle(x, c, f), (x', c', f')\rangle$, and itself involves assignment functions. We must therefore be careful to avoid circularity in the definition of $F^{\mathcal{M}}$, as is familiar from the work of Frank (1996) on modal subordination. Our problem and our solution are, however, slightly different than those for modal subordination. We will proceed in two steps. We will first define a set of bounded label-assignments, bounded in the sense that it only allows for a finite nesting of labels with a defined content into other labels. We give a semantics relative to such assignments of content to labels. The boundedness of assignments has an intuitive semantic counterpart: in any state, regardless of the linguistic actions executed so far, there is a finite bound n, for each agent i, on the maximal nesting of commitments $C_iC_i \ldots C_i\varphi$ that may evaluate to true for a non-tautological φ. In other words, there is always a maximal n such that agent's i commitments of order n are minimal (contain nothing besides logical tautologies). This is analogous to what happens in the "weak" dynamic commitment logic (without infinite sequences of acknowledgments) if we assume an initial model with minimal commitments. An immediate consequence is that it's impossible to reach common commitment on any content. We make this correspondence explicit below.

We will then extend our label-assignment functions to unbounded ones. Avoiding circularity in this case will require a quite tricky mathematical construction; however, the complexity of this construction will fade away thanks to two propositions stating that we can manipulate unbounded assignments exactly as bounded ones (namely, that a label assignment can evaluate any label into a set of state transitions, and that we can construct a new label-assignment from any label assignment σ, set of transitions T and label π by setting $\sigma(\pi) = T$). The semantics with unbounded assignments will be exactly the same as for bounded assignments.

In what follows, every definition is relative to a given model $\mathcal{M}$; in order to simplify notation we drop the $\cdot^{\mathcal{M}}$ when referring to the sets $F^{\mathcal{M}}, X^{\mathcal{M}}, S^{\mathcal{M}}$ and interpretations $[\![\cdot]\!]_0^{\mathcal{M}}$ and $[\![\cdot]\!]^{\mathcal{M}}$.

We define by simultaneous induction the sets F^n of rank-n label-assignments, and S^n of rank n states.

Definition 70. Rank of assignments

- $F^0 = \epsilon$, $S^0 = X \times C \times F^0$
- $F^{n+1} = (\wp(S^n \times S^n))^{\Pi}$ (functions from labels to set of transitions between rank-n states), $S^{n+1} = X \times C \times (\bigcup_{k \leq n} F^n)$
- $F = \bigcup_{n<\omega} F^n$, $S = X \times C \times F$

Informally, a label-assignment of rank n, is either the special symbol ϵ, or a function from discourse labels to a set of transitions (of rank $n-1$).

Our semantic requires only two operations on label-assignments: the evaluation of a label-assignment f at a given label π, and the substitution of a new set of transitions T for the set of transitions previously

[39] *i.e.* the label which is currently maximal w.r.t discourse subordination, this allows to dynamically exclude labels which introduce only intermediary contributions that do not constitute a commitment of their own, as *e.g.* antecedents to conditionals.

assigned to π yielding a new label-assignment denoted as $f[T/\pi]$. To avoid circularity, however, we cannot define substitution by any set of transitions T; it suffices for our purposes to restrict the definition to *bounded* sets of transitions, bounded in the sense that there must be some $k < \omega$ such that $T \subseteq S^k \times S^k$. $\Delta_B \subsetneq \wp(S \times S)$ is the set of such bounded sets of transitions ($\Delta_B = \bigcup_{n\in\omega} \wp(S^n \times S^n)$). The two operations (substitution and evaluation) are defined below.

Definition 71.
- $f(\pi)$ is already defined for $f \neq \epsilon$, since by construction, f is a function. We define in addition $\epsilon(\pi) = \bigcup_{x\in X, c\in C}\langle (x,c,\epsilon)(x,c,\epsilon)\rangle$ (this choice of definition makes ϵ represents a state with minimal commitments over logical tautologies only).
- for $T \in \Delta_B$, $\pi \in \Pi$, $f[T/\pi]$ is defined as the function f' such that:

$$\begin{cases} f'(\pi) = T \\ \forall \pi' \neq \pi\, f'(\pi') = f(\pi) \end{cases}$$

Note that, $T \in \Delta_B$ is sufficient to ensure that $f[T/\pi]$ stays in F.

Since the semantics depends on the considered set of relations and their semantics, we will assume that for every symbol $R \in \mathcal{R}$, assignment f and top labels c, a semantics $|R(\pi_1,\pi_2)|^f_c \in \Delta_B$ is defined, which, given an assignment to π_1 and π_2 provides a new proposition, that is to say, an element of Δ_B. $[\![\cdot]\!]$ interprets formulae into sets of transitions, and is defined inductively as follows (using infix notation $x\ \delta\ y$ as a syntactic sugar for $\langle x,y\rangle \in \delta$):

Definition 72 (Semantics for $\mathcal{L}_{\Pi,\mathcal{R}}$).
- For $\varphi \in \mathcal{L}_o$ $s\ [\![\varphi]\!]\ s'$ iff $s = (x,c,f)$, $s' = (x',c,f)$ and $x\ [\![\varphi]\!]_o\ x'$
- $(x,c,f)\ [\![R(\pi_1,\pi_2,\pi_3)]\!]\ (x',c',f')$ iff $c' = c[\pi_3/spk(\pi_3)]$ and $f' = f[(|R(\pi_1,\pi_2)|^f_c)/\pi_3]$
- $s\ [\![C_i\alpha]\!]\ s'$ iff $s = s' = \langle x,c,f\rangle$ and $\forall\langle s,s'\rangle \in f(c(i))\ \exists s''\ s'[\![\alpha]\!]s''$
- $(x,c,f)\ [\![\pi : \alpha]\!]\ (x',c',f')$ iff $f' = f[(f(\pi)\circ[\![\alpha]\!])/\pi]$
- $(x,c,f)\ [\![\pi_1 \overset{\pi}{\sim} \pi_2]\!]\ (x',c',f')$ iff $x = x'$ and $c' = c[\pi/spk(\pi)]$ and $f' = f[f(\pi_1)\cup f(\pi_2)/\pi]$
- $s\ [\![\neg\alpha]\!]\ s'$ iff $s = s'$ and there exists no $s'' \in S$ such that $s[\![\alpha]\!]s''$

With $c[\pi/i]$ defined as the unique c' such that $c'(i) = \pi$ and $\forall j \neq i, c'(j) = c(j)$

Since a bounded label-assignment f assigns a bounded set of transitions to any label, a easy induction on α show that $f(\pi)\circ[\![\alpha]\!]$ is always a bounded set of transitions as well, and thus $f[(f(\pi)\circ[\![\alpha]\!])/\pi]$ is always well defined and so is the semantics.

We provide in addition the semantics of two basic discourse relations: `Continuation` and `Ackn`:

Definition 73.

$$|\texttt{Continuation}(\pi_1,\pi_2)|^f_c = f(\pi_1)\circ f(\pi_2)$$
$$|\texttt{Ackn}(\pi_1,\pi_2)|^f_c = \{(\langle x,d,g\rangle,\langle y,e(spk(\pi_1)/\pi_1),h[f(\pi_1)/\pi_1]\rangle) \mid (\langle x,d,g\rangle,\langle y,e,h\rangle) \in f(c(spk(\pi_2)))\}$$

We can now link the propositional dynamic logic of the previous section and our dynamic semantics. We define a translation from $\mathcal{L}_D$ into $\mathcal{L}_{\Pi,\mathcal{R}}$ taking $\mathcal{L}_{\texttt{test}}$ as lower-level language. Let $M^o = \langle X, (R^o_i)_{i\in I}\rangle$ be the model of $\mathcal{L}_D$ whose set of world X is the set of valuations over signature PROP and $R^o_i = X \times X$ (thus M^o represents minimal commitments only to logical truths by each agent in I).

Algorithm 1 translation of $\mathcal{L}_D$ into $\mathcal{L}_{\Pi,\mathcal{R}}$

Assume a function $\texttt{fresh} : I \mapsto \Pi$ which enumerates each Π_i, *i.e.* such that each call $\texttt{fresh}(i)$ returns a label in Π_i and successive calls never return twice the same label.

function $t(\varphi, c)$ ▷ $\varphi \in \mathcal{L}_D \cup \mathcal{A}$, $c \in C$
  Case $\varphi = p \in \text{PROP}$:
    Return p
  Case $\varphi = C_i\varphi$:
    Return $C_i\, t(\varphi, c)$
  Case $\varphi \wedge \psi$:
    Return $t(\varphi, c) \wedge t(\psi, c)$
  Case $\varphi = [\alpha^i]\psi$:
    Let $\chi_\alpha = t(\alpha^i, c)$ and $\chi_\psi = t(\psi, c)$
    Return $\chi_\alpha \to \chi_\psi$
  Case $\varphi = \psi!^i$:
    Return $c(i) : t(\psi, c)$
  Case $\varphi = ack(\beta^x)^i$:
    Let $\pi_\beta = \texttt{fresh}(x)$
    Let $\chi_\beta = t(\beta^x, c[\pi_\beta/x])$
    Return $c(i) : \chi_\beta$
  Case $\varphi = (\alpha \sim \beta)^i$:
    Let $\pi_\alpha = fresh(i)$ and $\pi_\beta = fresh(i)$
    Let $\chi_\alpha = t(\alpha, c[\pi_\alpha/i])$ and $\chi_\beta = t(\beta, c[\pi_\beta/i])$
    Return $\chi_\alpha \wedge \chi_\beta \wedge \pi_\alpha \overset{c(i)}{\sim} \pi_\beta$
  end Case
end function

Proposition 15 (translation)**.** Algorithm 1, given as input a formula $\varphi \in \mathcal{L}_D$ and an initial "top" label $c \in C$ for each agent, yields a proposition $t(\varphi, c) = \chi \in \mathcal{L}_{\Pi,\mathcal{R}}$ such that

$$\langle M^\circ, x\rangle \vDash \varphi \text{ iff } \exists f \in F\ (x, c, \epsilon)\ [\![\chi]\!]\ (x, c, f)$$

Consider $\varphi = [(p! \sim q!)^i]([C_i p!^j]C_j\bot \wedge [ack(p!^j) \sim ack(q!^j)][(C_i p)!^j]\neg C_j\bot)$. This formula of $\mathcal{L}_D$ states that after i said something ambiguous between p and q, j is inconsistent in saying that i is committed to p, unless he first (ambiguously) acknowledges one of the two readings. φ is true in (any world of) $\mathcal{M}^\circ$. The translation of φ is the formula χ below (where symbols with a $\cdot^x$ exponent denote labels in Π^x):

$$(\pi_1^i : p \wedge \pi_2^i : q \wedge \pi_1^i \overset{\pi^i}{\sim} \pi_2^i) \to$$
$$\Big((\pi^j : C_i p) \to C_j\bot) \wedge \Big((\pi_1^j : \pi^i : p \wedge \pi_2^j : \pi^i : q \wedge \pi_1^j \overset{\pi^j}{\sim} \pi_2^j) \to ((\pi^j : C_i p) \to \neg C_j\bot)\Big)\Big)$$

χ first updates the content of labels π_1^i and π_2^i, set the content of π^i to be the union of the content of those two labels, then set π^i to be i's top label. Since this first state leaves the assignment to label π^j unchanged, If the initial assigment is ϵ, π^j contains no transition making $C_i p$ hold, hence executing the action $\pi^j : C_i p$ would empty π^j and commits j to the absurdum. The ambiguous action $(\pi_1^j : (\pi^i : p) \wedge \pi_2^j : (\pi^i : q) \wedge \pi_1^j \overset{\pi^j}{\sim} \pi_2^j)$ on the other hand upates π^j to contain at least one transition placing p in the content of π^i, which can then be selected through $\pi^j : C_i p$, and won't empties j's commitments. Thus, a state $\langle x, c, \epsilon\rangle$ is accepted by χ.

We now extend our semantics with assignments interpreting infinite descending chains of labels with defined content. This will enable us to do two things: 1) deal with infinite sequences of `Ackn` and 2) have

states in which agents are commonly committed to some non-tautological content. We build unbounded label-assignments over bounded ones. by exploiting a notion of bounded bisimulation over bounded label-assignments. In the following, $\bar{x}$ will generally denote a tuple $\langle x, c\rangle \in X \times C$.

Definition 74. We define n-bisimilarity ($\cong_n$) between bounded label-assignments inductively:

- $\forall f, f' \in F\ f \cong_0 f'$.
- $f_0 \cong_{n+1} f_1$ iff the two following conditions hold:

  (forth) $\forall \pi\, \forall \langle (\bar{x}, g_0), (\bar{x}', g_0')\rangle \in f(\pi) \exists \langle (\bar{x}, g_1)(\bar{x}', g_1')\rangle \in f'(\pi)$ such that $g_1 \cong_n g_1'$ and $g_0 \cong_n g_0'$.

  (back) $\forall \pi\, \forall \langle (\bar{x}, g_1), (\bar{x}', g_1')\rangle \in f'(\pi) \exists \langle (\bar{x}, g_0)(\bar{x}', g_0')\rangle \in f(\pi)$ such that $g_1 \cong_n g_1'$ and $g_0 \cong_n g_0'$.

To ease notational clutter, we extend the notation $\cong_n$ to pairs of transitions, by writing $\langle (\bar{x}, f), (\bar{y}, g)\rangle \cong_n \langle (\bar{x}', f'), (\bar{y}', g')\rangle$ as a shortcut for $\bar{x} = \bar{y}$, $\bar{x}' = \bar{y}'$ and $f \cong_n f'$ and $g \cong_n g'$. We are now ready to define unbounded assignments:

Definition 75 (Unbounded label-assignments)**.** The set of unbounded label-assignments F_ω is defined as the set of infinite sequences $\sigma = \langle \sigma_0, \sigma_1, \ldots, \sigma_i, \ldots\rangle$ of bounded label-assignments, such that: (i) $\forall i\ \sigma_i \cong_{i+1} \sigma_{i+1}$ and (ii) $\forall i\ \sigma_i \in F^{i+1}$.

The idea behind this construction is that, at each index, we find a label-assignment that is compatible with and refines the assignment of the preceding indices, allowing us to evaluate arbitrarily deep nestings of discourse labels by picking an assignment at a sufficiently high index.

Let $S_\omega = X \times C \times F_\omega$ denote the new set of states, now based on unbounded assignments (the analog to S). Corresponding transitions are thus in $S_\omega \times S_\omega$.

Definition 76. let $\sigma, \sigma' \in F_\omega$, $\sigma \cong \sigma'$ iff $\forall i\ \sigma_i \cong_{i+1} \sigma_i'$. $\cong$ is an equivalence relation.

We extend the notation $\cong$ to pairs of transitions, as we did for finite bissimulation. Finally, we introduce the notion of a diagonal, which will prove useful later on:

Definition 77 (Diagonal)**.** Let $\langle s_i = (\bar{x}, \sigma^i)\rangle_{i\in\omega}$ be a sequence of states with constant first components $\bar{x} = \langle x, c\rangle$, and such that $\sigma_0^0 \cong_1 \sigma_1^1 \cong_2 \sigma_2^2 \ldots \sigma_n^n \cong_{n+1} \sigma_{n+1}^{n+1}$ we call the diagonal of such a sequence, the state $\delta(\langle s_i\rangle_{i\in\omega}) = \langle \bar{x}, \langle \sigma_i^i\rangle_{i\in\omega}\rangle$

The last ingredients needed for our dynamic semantics with unbounded assignments are evaluation and substitution, which we now define:

Definition 78 (Evaluation)**.** Let $\sigma \in F_\omega$. Define $\sigma(\pi)$ as:

$$\sigma(\pi) = \{\langle (\bar{x}, \alpha), (\bar{y}, \beta)\rangle \mid \forall i \langle (\bar{x}, \alpha_i), (\bar{y}, \beta_i)\rangle \in \sigma_{i+1}(\pi) \text{ and } \alpha, \beta \in F^\omega\}$$

For any equivalence relation $\bowtie$, let $x = y[\bowtie]$ denote equality modulo $\bowtie$.

Proposition 16. if $\sigma \cong \sigma'$ then $\sigma(\pi) = \sigma'(\pi)[\cong]$

An unbounded assignment always assigns to a label a set of transitions closed under diagonal; i.e., whenever $\forall k \in \omega\ \langle s_k, s_k'\rangle \in \sigma(\pi)$, such that both $\delta_s = (\langle s_k\rangle_{k\in\omega})$ and $\delta_{s'} = \delta(\langle s_k'\rangle_{k\in\omega})$ are defined, we have $\langle (\delta_s, \delta_{s'})\rangle \in \sigma(\pi)$.

We now define the substitution operation. As in the bounded case, we must ensure that the set of transitions T we substitute is of the same kind that an unbounded label assignment can have as output. This time the restriction is not one of boundedness, but instead, following the previous remark, that T be closed under diagonal.

Definition 79 (Substitution)**.** Let $T \subseteq S_\omega \times S_\omega$ such that T is closed under diagonal. Define $T_i = \{\langle(\bar{x}, \alpha_i), (\bar{y}, \beta_i)\rangle \mid \langle(\bar{x}, \alpha), (\bar{y}, \beta)\rangle \in T\}$. For $\pi \in \Pi$ and $\sigma \in F_\omega$ define $\sigma[T/\pi]$ as

$$\sigma' = \sigma_0[T_0/\pi] \cdot \sigma_1[T_1/\pi] \cdot \ldots \sigma_n[T_n/\pi] \cdot \ldots$$

Note that since by definition of F_ω, $T_i \in S^{i+1} \times S^{i+1} \subseteq \Delta_B$, $\sigma_i[T_i/\pi]$ is well defined.

The final proposition we need is:

Proposition 17. $\sigma[T/\pi] \in F_\omega$. Furthermore $\sigma[T/\pi](\pi) = T[\cong]$.

Propositions 16 and 17 together imply that we can define our semantics exactly as we did for bounded assignment in definition definition 72, provided that the interpretation of discoures relations satisfies the following constraint: for every relation symbol R and assigment f, $|R(\pi_1, \pi_2)|^f$ has to be closed under diagonalization. This can be shown for our definition of `Continuation` and `Ackn`.

Diagonalization captures precisely the regularity required for an infinite sequence of actions to be representable as a simple model update, which, as we have seen, is the case for acknowledgments. The very reason that allowed us, in the previous section, to capture an infinite sequence of iterated acknowledgment in a finite action-structure, was that level n actions of the hierarchy of acknowledgments only affects commitments of order n. The model obtained after any acknowledgment of order 1 is 0-bisimilar to the initial model (actual facts are not modified). Applying second-order acknowledgments modifies only second order commitments, and the model after this second step is 1-bisimilar to the model at step one. Applying the successive level of the hierarchy of acknowledgment therefore yield a sequence of models such that the n^{th} is $n-$ bisimilar to the $(n+1)^{\text{th}}$. The reason why the finite action-structure semantically equivalent to iterated acknowledgments, is that it yields through update a single model, which is, for every integer n, $n-$bisimilar to the n^{th} model in the infinite sequence of updates. As the exact same relationship holds between an infinite sequence $\langle\sigma_k\rangle_{k\in\omega}$ of assignment which admits a diagonal and its diagonal, we can use infinite iterated acknowledgments in our dynamic semantics, and to that end we introduce a new relation $\texttt{Ackn}^\omega$:

Just as with the propositional infinite sequence of actions of section 10.1, we define first an infinite sequence of actions. Let $\pi \in \Pi, c \in C$ and for $i, j \in I$ let $\langle \pi^k_{ok^j_i} \rangle_{k\in\omega}$ be a sequence of pairwise disjoint labels in Π_i. Define level-n iterated acknowledgments of π, χ^{ack}_n, with $\chi_0 = \bigwedge_{i,j\in I} \texttt{Ackn}(\pi, \pi^0_{ok^j_i}, c(i))$ and $\forall n\ \chi_{n+1} = \chi_n \wedge \bigwedge_{i,j\in I} \texttt{Ackn}(c(j), \pi^k_{ok^j_i}, c(i))$. Let $x \in X$ and $f \in F$ and define $\langle x, c_n, f_n\rangle$ as the unique state such that $\langle x, c, f\rangle \, [\![\chi_n]\!] \, \langle x, c_n, f_n\rangle$. $\forall n\ c_n = c$, moreover the sequence $\langle x, c, f_n\rangle_{n\in\omega}$ admits a diagonal. Let $\delta^{ack}_{\langle x,c,f\rangle}$ denote this diagonal. We can finally add a construction $\texttt{Ackn}^\omega(\pi, (\pi^i)_{i\in I})$ to the language, which , first, copy each agent i's commitment into π^i (in order to not erase the content of the current top labels and keep track of previous states), and then to proceed to iterate acknowledgments. Define therefore for a pair $\langle c, f\rangle$, the pair $\bar{c} = c[(\pi^i/i)_{i\in I}]$ and $\bar{f} = f[(f(c(i))/\pi^i)_{i\in I}]$, this performs the "copy" of each agent's commitment into the new labels. Define:

$$s \, [\![\texttt{Ackn}^\omega(\pi, (\pi^i)_{i\in I})]\!] \, s' \text{ iff } s = \langle x, c, f\rangle \text{ and } s' = \delta^{ack}_{\langle x,\bar{c},\bar{f}\rangle}$$

10.4 Examples revisited

We will conclude this chapter with the relational version, using the language and interpretation we have set up in the previous section, of the solution our common commitment dilemma which we recall here: neither a systematic "strong" interpretation of dialog moves, nor a "weak" one is satisfactory, as it either yields meaningless acknowledging moves or impossibility of grounding. However, as we have made both

kind of interpretations part of a common semantic vocabulary, we are no longer commited to a systematic use of one or the other. The remaining question is thus: what is exactly the condition at which an agreement is reached that synchronous iterated acknowledgements indeed happened?

The answer we provided to this question is: one has to ask. Our solution builds on our treatment of ambiguity: acknowledgments, such as the one that 1 performs in (9.4.1-c) of example (9.4.1), are considered as ambiguous in their strength and only a confirming question and answer might raise this ambiguity, and reach a non-deniable common-commitment. Considering again 1's acknowledgment in (9.4.1-c), we represent 1's contribution (OK), as ambiguous between a simple acknowledgment by 1 and an acknowledgment of a common acknowledgment of 0's "no" answer. Representing 0's "no" as the propositional action $\neg\texttt{bank}!^{0}$, (9.4.1-c) is modeled in the propositional language as $ack(\texttt{bank}!^{1}) \sim ack(!^{\omega}_{0,1}(\texttt{bank}!^{0}))$.

In the relational semantics on the other hand, we have (9.4.1-b) as $\pi^{0} : \neg\texttt{bank}$, followed by (9.4.1-b)

$$\texttt{Ackn}^{\omega}(\pi^{0}, (\pi^{0}_{strong}, \pi^{1}_{strong})) \wedge \pi^{0} \overset{\pi'^{0}}{\sim} \pi^{0}_{strong} \wedge \texttt{Ackn}(\pi_{0}, \pi^{1}_{OK}, \pi^{1}_{weak}) \wedge \pi^{1}_{weak} \overset{\pi^{1}}{\sim} \pi^{1}_{strong}.$$

This move commits 1 (*via* π^{1}) to an ambiguous proposition. The confirming question following the acknowledgment in 9.4.1-b, is then modeled as asking 0 to raise the ambiguity. We did not discuss the semantics of questions, but questions in dynamics semantics have been discussed at length in the litterature (see *e.g.* Groenendijk (2003)). We could modifies states as to include issue partitions in order to represent questions, as we did in chapter 8. On the propositional side, the recent account in dynamic epistemic logic of van Benthem and Minica (2012) could also be integrated to the propositional dynamics of chapter 9. What is crucial independently of this choice, is that 1's commitments indeed licences a polar question $C^{*}C_{0}\neg bank?$, to which 0 answering yes brings a commitment $C_{0}C^{*}C_{0}\neg bank$ and, at thus, disambiguates 0's commitments in π'_{0} by selecting the content of π^{0}_{strong}. After such a move, whatever 0 may say, he cannot deny, at any level, that he commited to $\neg bank$. A simple acknowledgment of 0's answer yield the common commitment.

Our relational semantics also allows us to deal with self-corrections. Corrections need the full relational semantics, because one content is revised by another. Consider again example example (9.4.1), repeated here. We are interested in the correcting move by 1 in (10.4.1-d).

(10.4.1)a. 0: Did you have a bank account in this bank?
b. 1: No sir.
c. 0: OK. So you're saying that you did not have a bank account at Credit Suisse?
d. 1: No. sorry, in fact, I had an account there.
e. 0: OK thank you.

When 1 says *No. sorry, in fact, I had an account there*, the move attaches to 1's negative answer (10.4.1-b) to 0's initial question in (10.4.1-a). But the effects of this discourse move change the surrounding discourse structure. The semantics of Correction replaces the content of the original affirmative answer with the negative answer in the corrective move. Because of this, the follow up question of 0 and 1's second affirmative answer are now moot. To mirror Lascarides and Asher (2009), this revision requires that our states include a copy of the SDRS constructed for the dialogue up to reaching the present state, and that the revision is calculated using that structure, by 'rerunning' the whole sequence of moves, suitably updated: 1 and 0's commitments are recomputed on the revised SDRS. We thus model self-corrections as a revision of one's commitments.

Self-corrections thus erase the commitments of the corrected action and possibly also the commitments ensuing from subsequent dependent actions like its acknowledgment. An immediate consequence is that self-corrections make commitments, even common commitments, unstable (non-monotonic).

Conclusion

The work of the present thesis was driven by a general goal of providing a formal model of conversational meaning accounting for aspects of

1. **Logical consequences**, *i.e.*, what follows logically from what has been said, by either participant: the objective was to provide a formal answer to questions such as who said what? Do the participants agree on some given fact? does what A said entails a given proposition φ?
2. **Rationality**, *i.e.*, providing a model of what to say at a given point in time, given one's preferences and goals.
3. The interaction bewteen these, in particular regarding the status to give to implicatures and other pragmatic inferences as part of the meaning of an agent's contribution.
4. **The dynamics of conversational meaning**, *i.e.*, how to provide logical forms that adapt with time, in order to capture a formal account of meaning, at each step of the conversation; how to represent the dependance of a contribution on the context of previous conversational moves, and conversely, the impact of a given move on subsequent continuations.

We have addressed different aspects of these questions, in three different parts:

1. The first part of the thesis has been concerned with questions 1 and 3 of above. In this part, we have addressed differences in discourse structure representations, theoretical methods of comparison of representations, and investigated the foundations of a quantitative account of semantic similarity.
2. In the second part of the thesis we have addressed difficulties arising in strategic settings, regarding, mainly, questions 2 and 3. Driven by the idea that existing formal accounts of question 3 may lay the basis of a solution, we investigated a class of infinite games called Message Exchange games, involving a vocabulary equipped with a well-defined dynamic semantics, as a candidate solution.
3. In the third part, we have mostly focused on question 3, though we have established some link to credibility and grounding related to question 2. Our general object of investigation was the dynamics of public commitments in conversations, the semantics of acknowledgments, and the problem of grounding.

Contributions

- We first looked, in part I into the structure of the logical form of discourse and dialog: what structural constraints to adopt, what impact for a choice of constraints.

  To do this we adopted a semantically driven perspective: formalisms differ in the constraint they adopt, but they all agree that coherence relations have semantic consequences. Hence, their disagreement pertains as to whether these consequences can be directly read from the structure, or whether

the structure needs first to be interpreted, or transformed in some way. There is here an analogy with sentential syntax: each structural constituent certainly contributes to some part of the logical form, yet it is not always clear how and where exactly it must be integrated.

To address the above, we defined in chapter 3 a monadic second order language with two sorts, which we used to axiomatize each of the different formalisms' structural constraints, and when these constraints allow it, to switch from one representation to another.

The model-theory of this second order language allowed us to formalize the different kinds of interpretations that, we argued, are needed to extract from a given structure a logical form expressing semantic consequences transparently. In particular, we formalized different versions of the nuclearity principle of Rhetorical Structure Theory. This further pursues the analogy with sentential syntax: we have shown how some of these interpretations are indeed underspecified ones, which brings a novel understanding of the sort of information that theories of discourse structure encode.

This yields also a way to compare logical forms expressed over the same set of elementary discourse units.

As the more general problem of comparing logical forms, without restricting these forms to involve the same elementary units is also of interest, and, also, as logical forms undergo changes as conversations unfolds, we turned to the question of quantifying the impact of a new linguistic action. We adopted, in chapter 4, a more abstract setting and a more general notion of semantic similarity.

Assuming, for our foundational study, a view of conversations as simple as possible, *i.e.* sequences of atomic moves, equipped with an interpretation function taking them into a distinct, 'semantic' space, we investigated semantic metrics i.e., distances or similarity between sequences of linguistic moves that base the closeness of two conversations on that of their respective semantic interpretations.

We also formalized definitions of different levels of 'semanticness' for a (pseudo-)metric, and we drew a 'map' of the logical links between the axioms characterizing these definitions.

We then considered different possible structures for the semantic space (partial orders, general lattices, bounded lattices, set lattices) and defined some generic distances (assuming finite semantic universe though) for the different kind of spaces. We introduced, in particular, a distance based on measuring the possible continuations and showed that it generalizes (as does its link to the symmetric difference metric) as a distance defined in terms of shortest path in a finite bounded semi-lattice.

We have investigated a list of candidate axioms capturing different intuitions about links between the structure of the semantic space, the concatenation or internal properties of some sequences, and the semantic interpretation function. Most of the proposed axioms led to impossibility results. Yet some interesting facts remain, for instance, the continuation metric satisfies a non-trivial axiom linking the join operation in the lattice to the metric behaviour, which other distances violate.

- A second body of work, part II, focused on the interaction of logical forms and rationality, with a focus on strategic dialogs, where the interests of the participants diverge.

  We proposed in chapter 7 a game theoretic account of such conversations within a new perspective: conversation as infinite sequences of moves. An agent is successful if he plays certain sequences, otherwise he looses. The set of successful sequences forms her *winning condition*.

  These games bring a mathematical characterization of classes of conversational objectives describing the 'shape' that a successful conversation must take. Crucially, as we dispose of the complete, conversational unfolding, we can explain why a player adopts a given set of winning sequences on semantic grounds, avoiding both problems of locality, on the one hand, and trivializing non-cooperative behaviour through backward induction, on the other hand.

We introduced the notion of a Jury as an abstract third party monitoring conversations, and enforcing a form of 'good' behaviour from the participants. The Jury is a conceptual tool that allows to conceive of a form of communication in non-cooperative settings: intuitively, agents, even if they do not take their opponents' interests in consideration, are still committed to other forms of constraints (*e.g.* politeness, or simply, avoiding a public admission that they do not care for their opponents' interests).

Assuming a logical representation of the meaning of a sequence of moves (using, e.g., Segmented Discourse Representation Theory), we can formalize how the Jury enforces linguistic constraints which are generic necessary conditions on successful plays (staying coherent, consistent, credible). We can describe agents' preferences in term of the contents that agents commit to.

Finally, we established that ME games are determined games for Borel winning conditions, and that many intuitive winning conditions are Borel, therefore determined. We provided two theoretical explanations of a rational basis for one to engage ina determined losing conversations: misdirection, and conversational blindness.

- In part III, we established that the above framework requires a semantics expressive enough to express commitments about others' commitments.

  We therefore defined across chapters 8 to 10 a dynamic logic of (nested) commitments and integrated it in SDRT.

  We refined in chapter 8 the link between credibility and commitments.

  In chapters 9 and 10 we investigated the semantics of acknowledgments. We proposed two versions of our commitment dynamics: one where grounding is systematic and acknowledgments superfluous, and a weaker one, where acknowledgments are required but grounding, in terms of common commitment (as we argued seems to be what grounding demands) is not achievable in finite time. We proposed an ambiguity thesis for acknowledgment, mixing the strong and weak dynamics that would allow to reach common commitment *via* particular sequences of acknowledgment and confirmation questions.

  In chapter 10 we considered a link between the two dynamics, and a conservative embeding of the propositional dynamics of chapter 9 in a language with discourse relations similar to that of chapter 8.

Future Work

Discourse formalims

A first perspective for future work would be to apply our formalisms interlingua to problems of discourse parsing. RST, for instance, disposes of sets of annotated data across different corpora larger than those existing for SDRT or dependency structures. A possibility would thus be to use these data, and the partial correspondence we developed between RST and dependency trees to train a dependency discourse parser on RST data.

In the same order of thought, we could try using the similarity function defined in chapter 3 as a loss function. Of course this asks the question of the complexity of this function and requires to implement it efficiently. This similarity function essentially tries to align relations from one structure with those of another structure over the same set of EDUs, minimizing a global loss which is a function of each pair of aligned relations' differences in scope. Though we did not work this out, we conjecture that this can be implemented using the Hungarian algorithm. Computing the matrix of scores for each possible alignment of a relation of one strucure with a relation of the other structure should be doable in time $r \times e$ with r the number of relation in the structure that has the most relations, and e the number of EDUs. The Hungarian algorithm should then compute the result in time $O(r^3)$.

A second, both theoretical and practical problem that we wish to investigate is the possibility of using our second order formulation to achieve interesting tree-automata treatments, for instance for model checking, or a compact representation of one of our underspecified interpretations. This necessitates that we establish clearly what the tree-width of our classes of structures (especially for the less constrained one) is.

Semantic metrics

Most of our proposed axioms have led to trivialisation results. We think that the reason for this is that we have not structured the semantic space in the right way. The semantic space should involve some representationalism, *e.g.*, discourse graphs, but not too much. In particular it would be interesting to try and extract some relevant aspects of the model-theory of dynamic semantics in general and our dynamic propositional logic for commitments in particular, and pursue our study of semantic metrics on a semantic space and interpertation function reflecting these aspects.

On the long run, we would also like to examine whether we can find a theoretical application of a continuation-based distance in the settings of the ME games.

ME games

We have so far developed ME games as an abstract theoretical framework to account for strategic conversations in their generality. One thing we plan to do, is therefore considering small arena with a restricted vocabularies to test and refine the model.

As we have seen, the topological characterization of winning condition is lost when switching from ME to BM games. We therefore wish to investigate further conditions ensuring the existence of winning strategies, in particular in regards of the way winning conditions might be axiomatized using the commitment logic.

Public Commitment

We have relied on related, but different notions of ambiguity in chapter 8 and chapter 9 respectively. In chapter 8 we have been using dynamic disjunction, *i.e.*, non-determinacy. In the propositional settings this amounts to defining $\varphi! \sim \psi!$ as non determistically outputing a model updated with one or the other action: $\mathcal{M} \vDash [\varphi! \sim \psi!]\chi$ iff both updates of $\mathcal{M}$ by, respectively $[\varphi!]$ and $[\psi!]$, satisfy χ. In chapter 9 we have used another notion of ambiguity: $\varphi! \sim \psi!$, performed by agent i, duplicates the worlds in the commitment set of i, and applies a different actions to each of the copies. In fact we can show that both notion are linked. They coincide when evaluating formula without negated commitment, but do not behave in the same way w.r.t. negation. We plan to show that we can make the link explicit, by adding a second, $S4$ modality, representing ambiguity explicitely in the modal logic. Letting $[\sim]$ denote such modality, we plan to work out the dynamics for $[\sim]$ and recover for instance the notion of ambiguity of chapter 9, by showing that a proposition like $C_i\varphi$ in the semantics of chapter 9, is logically equivalently to proposition $[\sim]C_i\varphi$ in the refined semantics with an explicit modality for ambiguity. If this works, we would have $[\varphi! \sim \psi!]\chi$ in our propositional model iff $[\varphi! \sim \psi!][\sim]t(\chi)$ in the refined model, where $t(\chi)$ is obtained replacing $C_i\varphi$ with $[\sim]C_i\varphi$ everywhere. Moreover we think that the refined model would be ideal for a detailed account of clarification questions.

Finally, we did not propose a detailed, fully worked out treatment of rejections and corrections, which interacts importantly with commitments, and have a great strategic role too. We plan to extend our logic to a detailed treatment of rejections, and to further study how rejection moves should enter ME games' winning conditions, and what the propoerties of corresponding winning sets ensue.

Bibliography

Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-dac, M., Draoulec, A. L., Muller, P., Péry-Woodley, M., Preévot, L., Rebeyrolle, J., Tanguy, L., M.Vergez-Couret, and Vieu, L. (2012a). An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. In *Proceedings of LREC 2012*. ELRA.

Afantenos, S., Asher, N., Benamara, F., Cadilhac, A., Dégremont, C., Denis, P., Guhe, M., Keizer, S., Lascarides, A., Lemon, O., et al. (2012b). Developing a corpus of strategic conversation in the settlers of catan. In *SeineDial 2012-The 16th Workshop On The Semantics and Pragmatics Of Dialogue*.

Afantenos, S. D., Asher, N., Muller, P., Denis, P., and Danlos, L. (2010). Learning recursive segments for discourse parsing. In *Proceedings of LREC 2010*.

Alchourrón, C. E., Gärdenfors, P., and Makinson, D. (1985). On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530.

Allen, J. and Litman, D. (1987). A plan recognition model for subdialogues in conversations. *Cognitive Science*, 11(2):163–200.

Asher, N. (1986). Belief in discourse representation theory. *Journal of Philosophical Logic*, 15:127–189.

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.

Asher, N. (2013). Implicatures and discourse structure. *Lingua*, 132(0):13 – 28. SI: Implicature and Discourse Structure.

Asher, N. and Fernando, T. (1997). Nonincrementality and revision in dialogue. In *Proceedings of MunDial Workshop. University of Muenchen*.

Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.

Asher, N. and Lascarides, A. (2012). Strategic conversation. submitted.

Asher, N. and Lascarides, A. (2013). Strategic conversation. *Semantics and Pragmatics*, 6(2):2:1–:62.

Asher, N. and Mao, Y. (2000). Reasoning with negated defaults in commonsense entailment. In *Proceedings of the Logic Colloquium 2000*, Paris.

Asher, N. and Morreau, M. (1991). Commonsense entailment. In Mylopoulos, J. and Reiter, R., editors, *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 387–392, Los Altos, California. Morgan Kaufmann.

Asher, N. and Paul, S. (2013). Infinite games with uncertain moves. In *Proceedings of the First International Workshop on Strategic Reasoning*, pages 25–32, Rome.

Asher, N. and Pogodalla, S. (2011a). A montagovian treatment of modal subordination. In *Semantics and Linguistic Theory*, number 20, pages 387–405.

Asher, N. and Pogodalla, S. (2011b). SDRT and Continuation Semantics. In Onada, T., Bekki, D., and McCready, E., editors, *New Frontiers in Artificial Intelligence JSAI-isAI 2010 Workshops, LENLS, JURISIN, AMBN, ISS, Tokyo, Japan, November 18-19, 2010, Revised Selected Papers*, volume 6797 of *LNCS*, pages 3–15. Springer.

Asher, N. and Venant, A. (2016). Ok or not ok? commitments, acknowledgments and corrections. *Proceedings of Semantics and Linguistic Theory (SALT 25)*.

Asher, N., Venant, A., Muller, P., and Afantenos, S. D. (2011). Complex discourse units and their semantics. In *Contstraints in Discourse (CID 2011)*, Agay-Roches Rouges, France.

Aumann, R. and Hart, S. (2003). Long cheap talk. *Econometrica*, 71(6):1619–1660.

Aumann, R. J. and Maschler, M. (1995). *Repeated games with incomplete information*. MIT press.

Axelrod, R. M. (2006). *The evolution of cooperation*. Basic books.

Baldridge, J., Asher, N., and Hunter, J. (2007). Annotation for and Robust Parsing of Discourse Structure on Unrestricted Texts. *Zeitschrift für Sprachwissenschaft*, 26:213–239.

Baltag, A. and Moss, L. (2004). Logics for epistemic programs. *Synthese*, 139(2):165–224.

Baltag, A., Moss, L., and Solecki, S. (1999). The logic of public announcements, common knowledge and private suspicions. Technical Report SEN-R9922, Centrum voor Wiskunde en Informatica.

Baltag, A., Moss, L. S., and Solecki, S. (1998). The logic of public announcements, common knowledge, and private suspicions. In *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK '98, pages 43–56, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Bateman, J. and Rondhuis, K. J. (1997). Coherence relations : Towards a general specification. *Discourse Processes*, 24(1):3–49.

Benamara, F. and Taboada, M. (2015). Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152, Denver, Colorado. Association for Computational Linguistics.

Benz, A., Jäger, G., and van Rooij, R., editors (2005). *Game Theory and Pragmatics*. Palgrave Macmillan.

Black, E., Abney, S. P., Flickenger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J. L., Liberman, M., Marcus, M. P., Roukos, S., Santorini, B., and Strzalkowski, T. (1991). A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Speech and Natural Language, Proceedings of a Workshop held at Pacific Grove, California, USA, February 19-22. 1991*. Morgan Kaufmann.

Blackburn, P., Benthem, J. F. A. K. v., and Wolter, F. (2006). *Handbook of Modal Logic, Volume 3 (Studies in Logic and Practical Reasoning)*. Elsevier Science Inc., New York, NY, USA.

Blackburn, P., Gardent, C., and Meyer-Viol, W. (1993). Talking about trees. In *EACL 6*, pages 21–29.

Brown, P. and Levinson, S. (1978). *Politeness: Some Universals and Language Usage*. Cambridge University Press.

Cabrio, E., Tonelli, S., and Villata, S. (2013). A Natural Language Account for Argumentation Schemes. In *AI*IA - XIII Conference of the Italian Association for Artificial Intelligence - 2013*, Turin, Italie. Springer.

Carlson, L., Marcu, D., and Okurowski, M. (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, Denmark, September 2001.

Chang, C. and Keisler, H. J. (1973). *Model Theory*. North Holland Publishing.

Chaterjee, K. (2007). Concurrent games with tail objectives. *Theoretical Computer Science*, 388:181–198.

Church, A. (1940). A formulation of the simple theory of types. *The journal of symbolic logic*, 5(02):56–68.

Clark, H. (1996). *Using Language*. Cambridge University Press, Cambridge, England.

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In Resnick, L., Levine, J., and Teasley, S., editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association.

Cooper, R. (2005). Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2):99–112.

Courcelle, B. and Engelfriet, J. (2012). *Graph structure and monadic second-order logic: a language-theoretic approach*, volume 138. Cambridge University Press.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006). Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585.

Crammer, K. and Singer, Y. (2003). Ultraconservative online algorithms for multiclass problems. *J. Mach. Learn. Res.*, 3:951–991.

Crawford, V. and Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6):1431–1451.

Danlos, L. (2008). Strong generative capacity of RST, SDRT and discourse dependency DAGSs. In Benz, A. and Kuhnlein, P., editors, *Constraints in Discourse*, pages 69–95. Benjamins.

de Groote, P. (2006). Towards a montagovian account of dynamics. In *Semantics and Linguistic Theory*, pages 1–16.

DeVault, D. and Stone, M. (2007). Managing ambiguities across utterances in dialogue. In *Proceedings from the International Workshop on the Semantics and Pragmatics of Dialogue (DECALOG 2007)*, Trento, Italy.

Drew, P. (1992). Contested evidence in courtroom cross-examination: The case of a trial for rape. *Talk at work: Interaction in institutional settings*, pages 470–520.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and $< i > n < /i >$-person games. *Artificial intelligence*, 77(2):321–357.

duVerle, D. and Prendinger, H. (2009). A novel discourse parser based on support vector machine classification. In *Proceedings of ACL-IJCNLP 2009*, pages 665–673. ACL.

Egg, M. and Redeker, G. (2008). Underspecified discourse representation. *PRAGMATICS AND BEYOND NEW SERIES*, 172:117.

Egg, M. and Redeker, G. (2010). How Complex is Discourse Structure? In Calzolari, N., Choucri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of LREC'10*. ELRA.

Eiter, T. and Mannila, H. (1997). Distance measures for point sets and their computation. *Acta Informatica*, 34(2):109–133.

Farrell, J. (1993a). Meaning and credibility in cheap-talk games. *Games and Economic Behaviour*, 5:514–531.

Farrell, J. (1993b). Meaning and credibility in cheap-talk games. *Games and Economic Behavior*, 5(4):514–531.

Forbes, K., Miltsakaki, E., Prasad, R., Sarkar, A., Joshi, A., and Webber, B. (2001). D-LTAG System – discourse parsing with a lexicalized tree adjoining grammar. In *Proceedings of the ESSLLI-01 Workshop on Information Structure, Discourse Structure and Discourse Semantics*.

Frank, A. (1996). *Context dependence in modal constructions*. Ph.d. dissertation, IMS, University of Stuttgart, Stuttgart, Germany.

Franke, M. (2008). Meaning and inference in case of conflict. In Balogh, K., editor, *Proceedings of the 13th ESSLLI Student Session*, pages 65–74.

Franke, M. (2009). *Signal to Act: Game Theory in Pragmatics*. ILLC dissertation series. Institute for Logic, Language and Computation.

Franke, M., de Jager, T., and van Rooij, R. (2009). Relevance in cooperation and conflict. *Journal of Logic and Language*.

Frege, G. (1893). *Grundgesetze der Arithmetik*, volume I. Verlag Hermann Pohle, Jena.

Friedman, R. and Malone, P. (2010). *Rules of the Road: A Plaintiff Lawyers Guide to Proving Liability*. Trial Guides, 2nd edition.

Gale, D. and Stewart, F. M. (1953). Infinite games with perfect information. *Annals of Mathematical Studies*, 28:245–266.

Ginzburg, J. (2012). *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

Glazer, J. and Rubinstein, A. (2001). Debates and decisions: On a rationale of argumentation rules. *Games and Economic Behavior*, 36(2):158–173.

Glazer, J. and Rubinstein, A. (2004). On optimal rules of persuasion. *Econometrica*, 72(6):119–123.

Grädel, E. (2008). Banach-Mazur games on graphs. In Hariharan, R., Mukund, M., and Vinay, V., editors, *Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, pages 364–382.

Grice, H. P. (1967). *Studies in the Way of Words*, chapter Logic and Conversation, pages 22–40. Havard University Press, Cambridge, MA.

Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics Volume 3: Speech Acts*, pages 41–58. Academic Press.

Grice, H. P. (1989). *Studies in the Way of Words*. Harvard University Press, Cambridge, Massachusetts.

Groenendijk, J. (2003). Questions and answers: Semantics and logic. In *Proceedings of the 2nd CologNET-ElsET Symposium. Questions and Answers: Theoretical and Applied Perspectives*, pages 16–23.

Groenendijk, J. and Stokhof, M. (1984). *Studies on the Semantics of Questions and the Pragmatics of Answers*. PhD thesis, Centrale Interfaculteit, Amsterdam.

Groenendijk, J. and Stokhof, M. (1991). Dynamic predicate logic. *Linguistics and Philosophy*, 14:39–100.

Grosz, B. and Sidner, C. (1986). Attention, intentions and the structure of discourse. *Computational Linguistics*, 12:175–204.

Grosz, B. J. and Kraus, S. (1993). Collaborative plans for group activities. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 367–373, Los Altos, California. Morgan Kaufmann.

Hamblin, C. (1987). *Imperatives*. Blackwells.

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachussetts.

Hitzeman, J., Moens, M., and Grover, C. (1995). Algorithms for analyzing the temporal structure of discourse. In *Proceedings of the 7th Meeting of the European Chapter of the Association for Computational Linguistics*, pages 253–260.

Hobbs, J. (1979). Coherence and coreference. *Cognitive Science*, 3(1):67–90.

Hobbs, J. R. (1985). On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information, Stanford University.

Hobbs, J. R., Stickel, M., Appelt, D., and Martin, P. (1993a). Interpretation as abduction. *Artificial Intelligence*, 63(1–2):69–142.

Hobbs, J. R., Stickel, M., and Martin, P. (1993b). Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

Hunter, J., Asher, N., and Reese, B. (2006). Evidentiality and Intensionality: Two Uses of Reportative Constructions in Discourse. In Sidner, C., Harpur, J., Benz, A., and Kühnlein, P., editors, *Constraints in Discourse (CID), Maynooth, Ireland, 07/07/2006-09/07/2006*, pages 99–107, http://www.constraints-in-discourse.org/cid06/. National University of Ireland, Maynooth, Ireland.

Kamp, H. (1981). A theory of truth and semantic representation. In Groenendijk, J., Janssen, T., and Stokhof, M., editors, *Formal Methods in the Study of Language*, pages 277–322. Mathematisch Centrum, Amsterdam.

Kamp, H. (1990). Propositional attitudes. In Anderson, C. A. and Owens, J., editors, *The Role of Content in Logic, Language and Mind*. CSLI publications, University of Chicago Press.

Kamp, H. and Reyle, U. (1993). *From Discourse to the Lexicon: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers.

Kechris, A. (1995). *Classical descriptive set theory*. Springer-Verlag, New York.

Kehler, A. (2002). *Coherence, Reference and the Theory of Grammar*. CSLI Publications, Cambridge University Press.

Lamport, L. (1980). Sometime is sometimes not never: On the temporal logic of programs. In *Proceedings of the 7th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 174–185. ACM.

Lascarides, A. and Asher, N. (1993). Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16:437–493.

Lascarides, A. and Asher, N. (2009). Agreement, disputes and commitment in dialogue. *Journal of Semantics*, 26(2):109–158.

Lehmann, D., Magidor, M., and Schlechta, K. (2001). Distance semantics for belief revision. *Journal of Symbolic Logic*, 66(1):295–317.

Lewis, D. (1969). *Convention: A Philosophical Study*. Harvard University Press.

Libkin, L. (2004). *Elements of finite model theory*. Springer.

Lipman, B. and Seppi, D. (1995). Robust inference in communication games with partial provability. *Journal of Economic Theory*, 66:370–405.

Maier, E. (2010). Presupposing acquaintance: a unified semantics for de dicto, de re and de se belief reports. *Linguistics and Philosophy*, 32(5).

Mann, W. C. and Thompson, S. A. (1986). Rhetorical structure theory: Description and construction of text structures. In Kempen, G., editor, *Natural Language Generation: New Results in Artificial Intelligence*, pages 279–300.

Mann, W. C. and Thompson, S. A. (1987). Rhetorical structure theory: A framework for the analysis of texts. *International Pragmatics Association Papers in Pragmatics*, 1:79–105.

Marcu, D. (1996). Building up rhetorical structure trees. In *Proceedings of the thirteenth national conference on Artificial intelligence - Volume 2*, AAAI'96, pages 1069–1074. AAAI Press.

Martin, D. A. (1975). Borel determinacy. *Annals of Mathematics*, 102(2):363–371.

Mauldin, R., editor (1981). *The Scottish Book. Mathematics from the Scottish Café*. Birkhaüser.

Monjardet, B. (1981). Metrics on partially ordered sets — a survey. *Discrete Mathematics*, 35(1–3):173–184. Special Volume on Ordered Sets.

Montague, R. (1988). The proper treatment of quantification in ordinary english. In Kulas, J., Fetzer, J., and Rankin, T., editors, *Philosophy, Language, and Artificial Intelligence*, volume 2 of *Studies in Cognitive Systems*, pages 141–162. Springer Netherlands.

Muller, P., Afantenos, S., Denis, P., and Asher, N. (2012). Constrained decoding for text-level discourse parsing. In *COLING - 24th International Conference on Computational Linguistics*, Mumbai, Inde.

Muskens, R. (1996). Combining montague semantics and discourse representation. *Linguistics and Philosophy*, 19:143–186.

Oxtoby, J. (1957). The Banach-Mazur game and Banach category theorem. *Contribution to the Theory of Games*, 3:159–163.

Parikh, P. (1991). Communication and strategic inference. *Linguistics and Philosophy*, 14(5):473–514.

Parikh, P. (2000). Communication, meaning and interpretation. *Linguistics and Philosophy*, 25:185–212.

Parikh, P. (2001). *The Use of Language*. CSLI Publications, Stanford, California.

Polanyi, L. (1985). A theory of discourse structure and discourse coherence. In W. H. Eilfort, P. D. K. and Peterson, K. L., editors, *Papers from the General Session at the 21st Regional Meeting of the Chicago Linguistics Society*.

Polanyi, L. (1996). The linguistic structure of discourse. Technical Report CSLI-96-200, CSLI, Stanford University.

Polanyi, L., Culy, C., van den Berg, M., Thione, G. L., and Ahn, D. (2004). A rule based approach to discourse parsing. In *Proceedings of the 5th SIGDIAL Workshop in Discourse and Dialogue*, pages 108–117.

Polanyi, L. and Scha, R. (1984). A syntactic approach to discourse semantics. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING84)*, pages 413–419, Stanford.

Pollock, J. L. and Cruz, J. (1999). *Contemporary theories of knowledge*, volume 35. Rowman & Littlefield.

Popper, K. R. (1968). *Conjectures and refutations: The growth of scientific knowledge*. Harper & Row, New York.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The penn discourse treebank 2.0. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Purver, M. (2004). *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, Department of Computer Science, King's College, London.

Rabin, M. (1990). Communication between rational agents. *Journal of Economic Theory*, 51:144–170.

Roberts, C. (1989). Modal subordination and pronominal anaphora in discourse. *Linguistics and Philosophy*, 12(6):683–721.

Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5:6–1.

Roze, C. (2013). *Vers une algèbre des relations de discours*. PhD thesis, Université Paris-Diderot-Paris VII.

Sacks, H. (1992). *Lectures on Conversation*. Blackwell Publishers, Oxford. Edited by Gail Jefferson. This is the published version of lecture notes from 1967–1972.

Sagae, K. (2009). Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of IWPT'09*, pages 81–84. ACL.

Sampson, G. (2000). A proposal for improving the measurement of parse accuracy. *International Journal of Corpus Linguistics*, 5:53–68.

Sanders, T., Spooren, W., and Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse processes*, 15(1):1–35.

Searle, J. (1965). What is a speech act? In Black, M., editor, *Philosophy in America*, pages 615–628. Cornell University Press.

Searle, J. (1969). *Speech Acts*. Cambridge University Press.

Searle, J. R. (1962). Meaning and speech acts. *Philosophical Review*, 71:423–432.

Serre, O. (2004). Games with winning conditions of high borel complexity. In *Automata, Languages and Programming*, pages 1150–1162. Springer.

Solan, L. and Tiersma, P. (2005). *Speaking of Crime: The Language of Criminal Justice*. University of Chicago Press, Chicago, IL.

Spence, A. M. (1973). Job market signaling. *Journal of Economics*, 87(3):355Ð374.

Sperber, D. and Wilson, D. (1986). *Relevance*. Blackwells.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, 25(5):701–721.

Stede, M. (2004). The potsdam commentary corpus. In Webber, B. and Byron, D. K., editors, *ACL 2004 Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain. Association for Computational Linguistics.

Stone, M. and Lascarides, A. (2010). Grounding as implicature. In *Proceedings of the 14th SEMDIAL Workshop on the Semantics and Pragmatics of Dialogue*, pages 51–58, Poznan.

Subba, R. and Di Eugenio, B. (2007). Automatic discourse segmentation using neural networks. In *Proceedings of The 2007 Workshop on the Semantics and Pragmatics of Dialogue*, page 189.

Subba, R. and Di Eugenio, B. (2009). An effective discourse parser that uses rich linguistic information. In *Proceedings of HLT-NAACL*, pages 566–574. ACL.

Tarski, A. (1936). The concept of truth in formalized languages. In Tarski, A., editor, *Logic, Semantics, Metamathematics*, pages 152–278. Oxford University Press.

Traum, D. (1994). *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, Computer Science Department, University of Rochester.

Traum, D. (2008). Computational models of non-cooperative dialogue. In *Proceedings of the International Workshop on the Semantics and Pragmatics of Dialogue (LONDIAL)*, London.

Traum, D. and Allen, J. (1994). Discourse obligations in dialogue processing. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL94)*, pages 1–8, Las Cruces, New Mexico.

van Benthem, J. and Minica, S. (2012). Toward a dynamic logic of questions. *J. Philosophical Logic*, 41(4):633–669.

van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *Journal of Semantics*, 9(4):333–377.

van Rooij, R. (2003). Being polite is a handicap: towards a game theoretical analysis of polite linguistic behavior. In *Proceedings of Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 45–58.

van Rooij, R. (2004). Signalling games select horn strategies. *Linguistics and Philosophy*, 27:493–527.

Venant, A., Asher, N., and Dégremont, C. (2014). Credibility and its attacks. In *Proceedings of Semdial 2014*, Edinburgh, Scotland. Semdial.

Venant, Antoine & Asher, N. (2015). Dynamics of public commitments in dialogue. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 272–282, London, UK. Association for Computational Linguistics.

Vieu, L. (2011). On the semantics of discourse relations. In Asher, N. and Danlos, L., editors, *Constraints in Discourse (CID 2011), Agay-Roches Rouges, France*, page (electronic medium), http://www.inria.fr. INRIA.

Walton, D. N. (1984). *LOGICAL DIALOGUE-GAMES*. University Press of America, Lanham, Maryland.

Wolf, F. and Gibson, E. (2005a). Representing discourse coherence: A corpus-based analysis. *Computational Linguistics*, 31:249–288.

Wolf, F. and Gibson, E. (2005b). Representing discourse coherence: A corpus based study. *Computational Linguistics*, 31(2):249–287.

Zwick, U. and Paterson, M. S. (1995). The complexity of mean payoff games. In *Computing and Combinatorics*, pages 1–10. Springer.

Structures, commitments and games in strategic conversations

Résumé

Les effets d'une action linguistique dépendent du contexte. Cela pose plusieurs questions, auxquelles il est essentiel qu'un modèle linguistique réponde: comment représenter le contexte conversationnel, la façon dont celui-ci influe sur le "sens" de chaque contribution et sur le choix rationnel fait par un agent de ce qu'il va dire ensuite? Il existe plusieurs théories de la structure du discours, mais celles-ci ne s'accordent pas sur un ensemble précis de contraintes structurelles régissant le contexte conversationnel. Nous proposons un formalisme unifié pour traduire et comparer entre théories et représentations distinctes et nous étudions les fondations axiomatiques d'une mesure de déviation 'sémantique' entre deux contextes conversationnels, et de l'impact de l'entrée de nouveaux éléments dans le contexte sur cette deviation. Un second travail porte sur l'interaction entre forme logique et rationnalité dans les conversations, plus spécifiquement, lorsque les intérêts des participants divergent. Nous proposons un modèle en théorie des jeux, dans lequel une conversation est une séquence infinie de coups linguistiques. Dans ce cadre nous formalisons certaines contraintes linguistiques génériques comme des conditions nécessaires au succès d'un agent (rester consistant, cohérent, crédible). Les préférences des agents sont décrites par des contenus auxquels ceux-ci souhaitent, ou ne souhaitent pas s'engager. Crucialement, on peut justifier et expliquer via des considérations sémantiques le choix des objectifs conversationnels des agents, et montrer quand et comment certaines inferences (implicatures) survivent ou disparaissent. Cela nécessite une sémantique adéquate, pour l'obtenir nous définissons une logique modale dynamique des engagements publics. Celle-ci permet de représenter les déclaration des participants vis-à-vis de leur propre engagements et de ceux de leurs interlocuteurs. Cela permet enfin un modèle de "grounding" a granularité plus fine que les approches existantes, qui demeurent cependant axiomatisable comme des cas particulier.

Abstract

The effects of a linguistic action depend on its context of use. This raises a certain number of issues for a model of language use: how to represent the conversational context, its relation to the meanings that agents convey, and how to model agents' rational choice of the next thing to say, in context? There is, between existing theories of discourse structure, no general agreement on a precise set of structural constraints governing the conversational context. To remedy this, we propose a unified framewok to translate and compare between distinct theories and representations. We then lay the axiomatic foundations of metrics measuring the semantic deviation between two conversational contexts, and the changes brought into such a deviation, as new moves enter the context. A second body of work focuses on the modeling of conversational meaning and its interaction with that of rationnality in conversations, more specifically strategic dialogs, where the interest of the participants diverge. We propose a game theoretic account of such conversations, as infinite sequences of linguistic moves. We formalize linguistic constraints that are generic necessary conditions on successful plays (staying coherent, consistent, credible), and describe agents' preferences in terms of the contents that agents commit to. Crucially, we can describe a player's objective and explain why it is adopted on semantic grounds. We show on this basis how and when inferences to non-litteral meaning survives or are cancelled. As this requires a semantics expressive enough, we define a dynamic logic of public commitments to represent participants' commitments about the content of theirs, or their opponent's moves, and keep those representations subject to a sound notion of logical consequence (and hence, of consistency). This yields an account of acknowledgment and grounding more formal and fine-grained than traditional approaches, recoverable as particular cases.