Chapter 1
Models of Persuasion Dialogue

Henry Prakken

1.1 Introduction

This chapter[1] reviews formal dialogue systems for persuasion. In persuasion dialogues two or more participants try to resolve a conflict of opinion, each trying to persuade the other participants to adopt their point of view. Dialogue systems for persuasion regulate how such dialogues can be conducted and what their outcome is. Good dialogue systems ensure that conflicts of view can be resolved in a fair and effective way (6). The term 'persuasion dialogue' was coined by Walton (13) as part of his influential classification of dialogues into six types according to their goal. While *persuasion* aims to resolve a difference of opinion, *negotiation* tries to resolve a conflict of interest by reaching a deal, *information seeking* aims at transferring information, *deliberation*deliberation wants to reach a decision on a course of action, *inquiry* is aimed at "growth of knowledge and agreement" and *quarrel* is the verbal substitute of a fight. This classification leaves room for shifts of dialogues of one type to another. In particular, other types of dialogues can shift to persuasion when a conflict of opinion arises. For example, in information-seeking a conflict of opinion could arise on the credibility of a source of information, in deliberation the participants may disagree about likely effects of plans or actions and in negotiation they may disagree about the reasons why a proposal is in one's interest.

The formal study of dialogue systems for persuasion was initiated by Hamblin (5). Initially, the topic was studied only within philosophical logic and argumentation theory (15; 7), but later several fields of computer science also became interested in this topic. In general AI the embedding of nonmonotonic logic in models of persuasion dialogue was seen as a way to deal with resource-bounded reasoning (6; 2), while in AI & Law persuasion was seen as an appropriate model of legal

Henry Prakken
Department of Information and Computing Sciences, Utrecht University, and Faculty of Law, University of Groningen, e-mail: henry@cs.uu.nl

[1] This chapter is a revised and updated version of (11).

procedures (4). In intelligent tutoring, systems for teaching argumentation skills have been founded on models of persuasion dialogue (16). Finally, in the field of multi-agent systems dialogue systems have been incorporated into models of rational agent interaction (8).

To delineate the scope of this chapter, it is useful to discuss what is the subject matter of dialogue systems. According to Carlson (3) dialogue systems define the principles of coherent dialogue, that is, the conditions under which an utterance is appropriate. The leading principle here is that an utterance is appropriate if it furthers the goal of the dialogue. For persuasion this means that an utterance should contribute to the resolution of the conflict of opinion that triggered the persuasion. Thus according to Carlson the principles governing the use of utterances should not be defined at the level of individual speech acts but at the level of the dialogue in which the utterance is made. Carlson therefore proposes a game-theoretic approach to dialogues, in which speech acts are viewed as moves in a game and rules for their appropriateness are formulated as rules of the game. Most work on formal dialogue systems for persuasion follows this approach and therefore this chapter will assume a game format of dialogue systems. It should be noted that the term *dialogue system* as used in this chapter only covers the rules of the game, i.e., which moves are allowed; it does not cover principles for playing the game well, i.e., strategies and heuristics for the individual players. The latter are instead aspects of agent models.

Below in Section 1.2 an example persuasion dialogue will be presented, which will be used for illustration throughout the paper. Then in Section 1.3 a formal framework for specifying dialogue game systems is proposed, which in Section 1.4 is instantiated for persuasion dialogues and in Section 1.5 is used for discussing and comparing three systems proposed in the literature.

1.2 An example persuasion dialogue

The following example persuasion dialogue exhibits some typical features of persuasion and will be used in this chapter to illustrate different degrees of expressiveness and strictness of the various persuasion systems.

`Paul`: My car is safe. (*making a claim*)
`Olga`: Why is your car safe? (*asking grounds for a claim*)
`Paul`: Since it has an airbag, (*offering grounds for a claim*)
`Olga`: That is true, (*conceding a claim*) but this does not make your car safe. (*stating a counterclaim*)
`Paul`: Why does that not make my care safe? (*asking grounds for a claim*)
`Olga`: Since the newspapers recently reported on airbags expanding without cause. (*stating a counterargument by providing grounds for the counterclaim*)
`Paul`: Yes, that is what the newspapers say (*conceding a claim*) but that does not prove anything, since newspaper reports are very unreliable sources of technological information. (*undercutting a counterargument*)

`Olga`: Still your car is still not safe, since its maximum speed is very high. (*alternative counterargument*)
`Paul`: OK, I was wrong that my car is safe.

This dialogue illustrates several features of persuasion dialogues.

- Participants in a persuasion dialogue not only exchange arguments and counterarguments but also express various propositional attitudes, such as claiming, challenging, conceding or retracting a proposition.
- As for arguments and counterarguments it illustrates the following features.
  - An argument is sometimes attacked by constructing an argument for the opposite conclusion (as in Olga's two counterarguments) but sometimes by saying that in the given circumstances the premises of the argument do not support its conclusion (as in Paul's counterargument). This is Pollock's well-known distinction between rebutting and undercutting counterarguments (9).
  - Counterarguments are sometimes stated at once (as in Paul's undercutter and Olga's last move) and are sometimes introduced by making a counterclaim (as in Olga's second and third move).
  - Natural-language arguments sometimes leave elements implicit. For example, Paul's second move arguably leaves a commonsense generalisation 'Cars with airbags usually are safe' implicit.
- As for the structure of dialogues, the example illustrates the following features.
  - The participants may return to earlier choices and move alternative replies: in her last move Olga states an alternative counterargument after she sees that Paul had a strong counterattack on her first counterargument. Note that she could also have moved the alternative counterargument immediately after her first, to leave Paul with two attacks to counter.
  - The participants may postpone their replies, sometimes even indefinitely: with her second argument why Paul's car is not safe, Olga postpones her reply to Paul's counterattack on her first argument for this claim; if Paul fails to successfully attack her second argument, such a reply might become superfluous.

1.3 Elements of dialogue systems

In this section a formal framework for specifying dialogue systems is proposed. To summarise, dialogue systems have a *dialogue goal* dialogue goal and at least two *participants*, who can have various *roles*. Dialogue systems have two languages, a a *communication language* wrapped around a *topic language*. Sometimes, dialogues take place in a *context* of fixed and undisputable knowledge, such as the relevant laws in a legal dispute. The heart of a dialogue system is formed by a *protocol*, specifying the allowed moves at each point in a dialogue, the *effect rules*, specifying the effects of utterances on the participants' commitments, and the *outcome rules*,

defining the outcome of a dialogue. Two kinds of protocol rules are sometimes separately defined, viz. *turntaking* and *termination* rules.

Let us now specify these elements more formally. The definitions below of dialogues, protocols and strategies are based on Chapter 12 of (1) as adapted in (10). As for notation, the complement $\overline{\varphi}$ of a formula φ is $\neg\varphi$ if φ is a positive formula and ψ if φ is a negative formula $\neg\psi$.

Definition 1. (Dialogue systems) A *dialogue system* consists of the following elements.

- A *topic language* $\mathscr{L}_t$, closed under classical negation.
- A *communication language* $\mathscr{L}_c$, consisting of a set of *speech acts* with a *content*. The set of *dialogues*, indexdialogues denoted by $M^{\leq\infty}$, is the set of all sequences from $\mathscr{L}_c$, and the set of *finite dialogues*, denoted by $M^{<\infty}$, is the set of all finite sequences from $\mathscr{L}_c$. For any dialogue $d = m_1,\ldots,m_n,\ldots$, the subsequence $m_1,\ldots,m_i$ is denoted with d_i.
- A *dialogue purpose*.
- A set $\mathscr{A}$ of *participants* (or 'players') and a set $\mathscr{R}$ of *roles*, defined as disjoint subsets of $\mathscr{A}$. A participant a may or may not have a, possibly inconsistent, *belief base* $\Sigma_a \subseteq Pow(\mathscr{L}_t)$, which may or may not change during a dialogue. Furthermore, each participant has a, possibly empty set of *commitments* $C_a \subseteq \mathscr{L}_t$, which usually changes during a dialogue.
- A *context* $K \subseteq \mathscr{L}_t$, containing the knowledge that is presupposed and must be respected during a dialogue. The context is assumed consistent and remains the same throughout a dialogue.
- A *logic* L for $\mathscr{L}_t$, which may or may not be monotonic and which may or may not be argument-based.
- A set of *effect rules* C for $\mathscr{L}_c$, specifying for each utterance $\varphi \in \mathscr{L}_c$ its effects on the commitments of the participants. These rules are specified as functions
  - $C_a : M^{<\infty} \longrightarrow Pow(\mathscr{L}_t)$

  Changes in commitments are completely determined by the last move in a dialogue and the commitments just before making that move:
  - If $d = d'$ then $C_a(d,m) = C_a(d',m)$
- A *protocol* Pr for $\mathscr{L}_c$, specifying the allowed (or 'legal') moves at each stage of a dialogue. Formally, A *protocol* on $\mathscr{L}_c$ is a function Pr with domain the context plus a nonempty subset D of $M^{<\infty}$ taking subsets of $\mathscr{L}_c$ as values. That is:
  - $Pr : Pow(\mathscr{L}_t) \times D \longrightarrow Pow(\mathscr{L}_c)$

  such that $D \subseteq M^{<\infty}$. The elements of D are called the *legal finite dialogues*. The elements of $Pr(d)$ are called the moves allowed after d. If d is a legal dialogue and $Pr(d) = \varnothing$, then d is said to be a *terminated* dialogue. Pr must satisfy the following condition: for all finite dialogues d and moves m, $d \in D$ and $m \in Pr(d)$ iff $d,m \in D$.

It is useful (although not strictly necessary) to explicitly distinguish elements of a protocol that regulate turntaking and termination:

- A *turntaking* function is a function $T : D \times Pow(\mathscr{L}_t) \longrightarrow Pow(\mathscr{A})$. A *turn* of a dialogue is defined as a maximal sequence of moves in the dialogue in which the same player is to move. Note that T can designate more than one player as to-move next.
- *Termination* is above defined as the case where no move is legal. Accordingly, an explicit definition of termination should specify the conditions under which Pr returns the empty set.

- *Outcome rules* O^K, defining the outcome of a dialogue given a context. For instance, in negotiation the outcome is an allocation of resources, in deliberation it is a decision on a course of action, and in persuasion dialogue it is a winner and a loser of the persuasion dialogue. The outcome must be defined for terminated dialogues and may be defined for nonterminated ones; in the latter case the outcome rules capture an 'anytime' outcome notion.

Note that no relations are assumed between a participant's commitments and belief base. Commitments are an agent's publicly declared points of view about a proposition, which need not coincide with the agent's internal beliefs.

Definition 2. (Some protocol types)

- A protocol has a *public semantics* if the set of legal moves is always independent from the agents' belief bases.
- A protocol is *context-independent* if the set of legal moves and the outcome is always independent of the context, so if $Pr(K,d) = Pr(\varnothing,d)$ and $O^K(d) = O^{\varnothing}(d)$ for all K and d.
- A protocol Pr is *fully deterministic* if Pr always returns a singleton or the empty set. It is *deterministic in* $\mathscr{L}_c$ if the set of moves returned by Pr at most differ in their content but not in their speech act type.
- A protocol is *unique-move* if the turn shifts after each move; it is *multiple-move* otherwise.

Paul and Olga (ct'd): The protocol in our running example is multiple-move.

Dialogue participants can have strategies and heuristics for playing the dialogue game in ways that promote their individual dialogue goal. The notion of a *strategy* for a participant a can be defined in the game-theoretical sense, as a function from the set of all finite legal dialogues in which a is to move into $\mathscr{L}_c$. A strategy for a is a *winning strategy* if in every dialogue played in accord with the strategy a realises his dialogue goal (for instance, winning in persuasion). *Heuristics* generalise strategies in two ways: they may leave the choice for some dialogues undefined and they may specify more than one move as a choice option. More formally:

Definition 3. (strategies and heuristics) Let D_a, a subset of D, be the set of all dialogues where a is to move, and let D'_a be a subset of D_a. Then a strategy and a heuristic for a are defined as functions s_a and h_a as follows.

- $s_a : D_a \longrightarrow \mathscr{L}_c$
- $h_a : D'_a \longrightarrow Pow(\mathscr{L}_c)$

1.4 Persuasion

Let us now become more precise about persuasion. Walton & Krabbe (14) define persuasion dialogues as dialogues with as goal to resolve a conflict of points of view between at least two participants. A *point of view* with respect to a proposition can be positive (for), negative (against) or doubtful. The participants aim to persuade the other participant(s) to accept their point of view. According to Walton & Krabbe a conflict is resolved if all parties share the same point of view on the proposition that is at issue. They distinguish *disputes* as a subtype of persuasion dialogues where two parties disagree about a single proposition φ, such that at the start of the dialogue one party has a positive (φ) and the other party a negative ($\neg\varphi$) point of view towards the proposition.

Dialogue systems for persuasion can be formally defined as a particular class of instantiations of the general framework.

Definition 4. (dialogue systems for persuasion) A *dialogue system for persuasion* is a dialogue system with at least the following instantiations of Definition 1.

- The *dialogue purpose* is resolution of a conflict of opinion about one or more propositions, called the *topics* $T \subseteq \mathscr{L}_t$. This dialogue purpose gives rise to the following participant roles and outcome rules.
- The participants can have the following *roles*. To start with, $prop(t) \subseteq \mathscr{A}$, the *proponents* of topic t, is the (nonempty) set of all participants with a positive point of view towards t. Likewise, $opp(t) \subseteq \mathscr{A}$, the *opponents* of t, is the (nonempty) set of all participants with a doubtful point of view toward a topic t. Together, the proponents and opponents of t are called the *adversaries* with respect to t. For any t, the sets $prop(t)$ and $opp(t)$ are disjoint but do not necessarily jointly exhaust $\mathscr{A}$. The remaining participants, if any, are the *third parties* with respect to t, assumed to be neutral towards t.
  Note that this allows that a participant is a proponent of both t and $\neg t$ or has a positive attitude towards t and a doubtful attitude towards a topic t' that is logically equivalent to t. Since protocols can deal with such situations in various ways, this should not be excluded by definition.
- The *Outcome rules* of systems for persuasion dialogues define for a dialogue d, context K and topic t the *winners* and *losers* of d with respect to topic d. More precisely, O consists of two partial functions w and l:
  - $w : D \times Pow(\mathscr{L}_t) \times \mathscr{L}_t \longrightarrow Pow(\mathscr{A})$
  - $l : D \times Pow(\mathscr{L}_t) \times \mathscr{L}_t \longrightarrow Pow(\mathscr{A})$

  such that they are defined at least for all terminated dialogues but only for those t that are a topic of d. These functions will be written as $w_t^K(d)$ and $l_t^K(d)$ or,

if there is no danger for confusion, as $w_t(d)$ and $l_t(d)$. They further satisfy the following conditions for arbitrary but fixed context K:

- $w_t(d) \cap l_t(d) = \varnothing$
- $w_t(d) = \varnothing$ iff $l_t(d) = \varnothing$
- if $|\mathscr{A}| = 2$, then $w_t(d)$ and $l_t(d)$ are at most singletons

- Next, to make sense of the notions of proponent and opponent, their commitments at the start of a dialogue should not conflict with their points of view.
  - If $a \in prop(t)$ then $\bar{t} \notin C_a(\varnothing)$
  - If $a \in opp(t)$ then $t \notin C_a(\varnothing)$
- Finally, in persuasion at most one side in a dialogue gives up, i.e.,
  - $w_t(d) \subseteq prop(t)$ or $w_t(d) \subseteq opp(t)$; and
  - If $a \in w_t(d)$ then
    - if $a \in prop(t)$ then $t \in C_a(d)$
    - if $a \in opp(t)$ then $t \notin C_a(d)$

  These conditions ensure that a winner did not change its point of view. Note that they make that two-person persuasion dialogues are zero-sum games. Perhaps this is the main feature that sets persuasion apart from information seeking, deliberation and inquiry.

Note that the two last winning conditions of the last bullet lack their only-if part. This is to allow for a distinction between so-called *pure persuasion* and *conflict resolution*. indexconflict resolution The outcome of pure persuasion dialogues is fully determined by the participants' points of view and commitments:

Definition 5. (types of persuasion systems)

- A dialogue system is for *pure persuasion* iff for any terminated dialogue d it holds that $a \in w_t(d)$ iff
  - either $a \in prop(t)$ and $t \in C_{a'}(d)$ for all $a' \in prop(d) \cup opp(d)$
  - or $a \in opp(t)$ and $t \notin C_{a'}(d)$ for all $a' \in prop(d) \cup opp(d)$
- Otherwise, it is for *conflict resolution*.

In addition, pure persuasion dialogues are assumed to terminate as soon as the right-hand-side conjuncts of one of these two winning conditions hold.

Paul and Olga (ct'd): In our running example, if the dialogue is regulated by a protocol for pure persuasion, it terminates after Paul's retraction.

In conflict resolution dialogues the outcome is not fully determined by the participant's points of view and commitments. In other words, in such dialogues it is possible that, for instance, a proponent of φ loses the dialogue about φ even if at termination he is still committed to φ. A typical example is legal procedure, where a third party can determine the outcome of the case. For instance, a crime suspect can be convicted even if he maintains his innocence throughout the case.

If the system has an anytime outcome notion, then another distinction can be made (6): a protocol is *immediate-response* if the turn shifts just in case the speaker is the 'current' winner and if it then shifts to a 'current' loser.

As for the communication language and effect rules, some common elements can be found throughout the literature. Below are the most common speech acts, with their informal meaning and the various names they have been given in the literature.[2]

- *claim* φ (assert, statement, ...). The speaker asserts that φ is the case.
- *why* φ (challenge, deny, question, ...) The speaker challenges that φ is the case and asks for reasons why it would be the case.
- *concede* φ (accept, admit, ...). The speaker admits that φ is the case.
- *retract* φ (withdraw, no commitment, ..) The speaker declares that he is not committed (any more) to φ. Retractions are 'really' retractions if the speaker is committed to the retracted proposition, otherwise it is a mere declaration of non-commitment (for example, in reply to a question).
- φ *since* S (argue, argument, ...) The speaker provides reasons why φ is the case. Some protocols do not have this move but instead require that reasons be provided by a *claim* φ or *claim* S move in reply to a *why* ψ move (where S is a set of propositions). Also, in some systems the reasons provided for φ can have structure, for example, of a proof three or a deduction.
- *question* φ The speaker asks the hearers' opinion on whether φ is the case.

Paul and Olga (ct'd): In this communication language our example from Section 1.2 can be more formally displayed as follows:

P_1: *claim* safe	O_2: *why* safe
P_3: safe *since* airbag	O_4: *concede* airbag
	O_5: *claim* $\neg$ safe
P_6: *why* $\neg$ safe	O_7: $\neg$ safe *since* newspaper
P_8: *concede* newspaper	
P_9: so what *since* $\neg$ newspapers reliable	O_{10}: $\neg$ safe *since* high max. speed
P_{11}: *retract* safe	

Most dialogue systems have a notion of typical replies to certain speech acts, although usually this is left implicit in the replies that are allowed by the protocol rules. In most systems these typical replies are as displayed in Table 1.1.

Paul and Olga (ct'd): With this table our running example can be displayed as in Figure 1.1, where the boxes stand for moves and the links for reply relations.

The reply notion induces another distinction between dialogue protocols.

Definition 6. A dialogue protocol is *unique-reply* if at most one reply to a move is allowed throughout a dialogue; otherwise it is *multiple-reply*.

Paul and Olga (ct'd): The protocol governing our running example is multiple-reply, as illustrated by the various branches in Figure 1.1.

[2] To make this chapter more uniform, the present terminology will be used even if the original publication of a system uses different terms.

Table 1.1 Locutions and typical replies

Locutions	Replies
claim φ	*why* φ, *claim* $\overline{\varphi}$, *concede* φ
why φ	φ *since* S (alternatively: *claim* S), *retract* φ
concede φ	
retract φ	
φ *since* S	*why* ψ ($\psi \in S$), *concede* ψ ($\psi \in S$), φ' *since* S
question φ	*claim* φ, *claim* $\overline{\varphi}$, *retract* φ

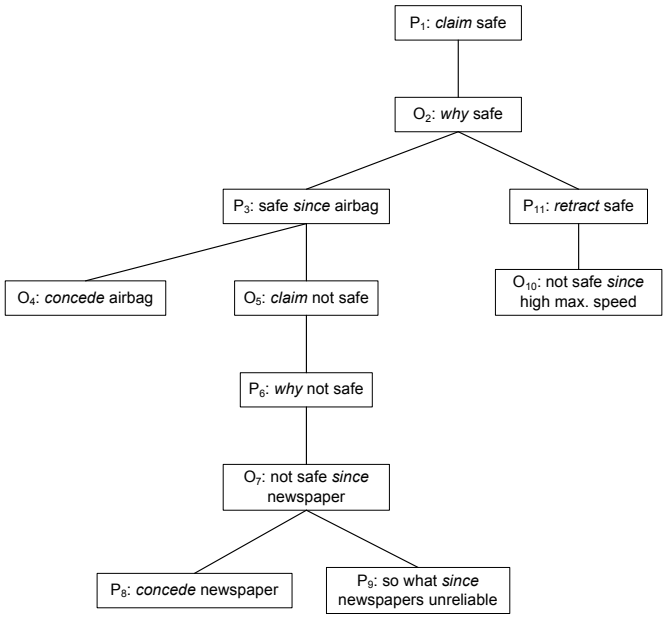

Fig. 1.1 Reply structure of the example dialogue

As for the commitment rules, the following ones are generally accepted in the literature. (Below *pl* denotes the speaker of the move; effects on the other parties' commitments are only specified when a change is effected.)

- If $pl(m) = claim(\varphi)$ then $C_{pl}(d,m) = C_{pl}(d) \cup \{\varphi\}$
- If $pl(m) = why(\varphi)$ then $C_{pl}(d,m) = C_{pl}(d)$
- If $pl(m) = concede(\varphi)$ then $C_{pl}(d,m) = C_{pl}(d) \cup \{\varphi\}$
- If $pl(m) = retract(\varphi)$ then $C_{pl}(d,m) = C_{pl}(d) - \{\varphi\}$
- If $pl(m) = \varphi$ *since* S then $C_{pl}(d,m) \supseteq C_{pl}(d) \cup prem(A)$

The rule for *since* uses $\supseteq$ since such a move may commit to more than just the premises of the moved argument. For instance, in (10) the move also commits to φ, since arguments can also be moved as counterarguments instead of as replies to

challenges of a claim. And in some systems that allow incomplete arguments, such as (14), the move also commits the speaker to the material implication $S \rightarrow \varphi$.

Paul and Olga (ct'd): According to these rules, the commitment sets of Paul and Olga at the end of the example dialogue are

- $C_P(d_{11}) \supseteq \{$airbag, newspaper, $\neg$ newspapers reliable$\}$
- $C_O(d_{11}) \supseteq \{\neg$ safe, airbag, newspaper, high max. speed$\}$

1.5 Three systems

Now three persuasion protocols from the literature will be discussed. The first is primarily based on commitments, the second defines protocols as finite state machines, while the third exploits an explicit reply structure on the communication language.

1.5.1 Walton and Krabbe (1995)

The first system to be discussed is Walton & Krabbe's dialogue system PPD for "permissive persuasion dialogues" (14). In PPD, dialogues have no context. The players are called White (W) and Black (B). They are assumed to declare zero or more "assertions" and "concessions" in an implicit preparatory phase of a dialogue. Each participant is proponent of his own and opponent of the other participant's initial assertions. B must have declared at least one assertion, and W starts a dialogue. The communication language consists of challenges, (tree-structured) arguments, concessions, questions, resolution demands ("resolve"), and two retraction locutions, one for assertion-type and one for concession-type commitments. It has no explicit reply structure but the protocol reflects the reply structure of Table 1.1 above.

The logical language is that of propositional logic and the logic consists of an incomplete set of deductively valid inference rules: they are incomplete to reflect that for natural language no complete logic exists. Although an argument may thus be incomplete, its mover becomes committed to the material implication *premises* $\rightarrow$ *conclusion*, which is then open for discussion.

The commitment rules are standard but Walton & Krabbe distinguish between several kinds of commitments for each participant, viz. *assertions*, *concessions* and *dark-side* commitments. Initial assertions and premises of arguments are placed in the assertions while conceded propositions are placed in the concessions. Only assertions can be challenged. Dark-side commitments are hidden or veiled commitments of an agent, of which they are often unaware. This makes them hard to model computationally, for which reason they will be ignored below.

The protocol is driven by two main factors: the contents of the commitment sets and the content of the last turn. W starts and in their first turn both W and B either concede or challenge each initial assertion of the other party. Then each turn must

reply to all moves in the other player's last turn except concessions and retractions; in particular, for *since* moves each premise must be conceded or challenged, including the hidden premise of incomplete arguments. Multiple replies are allowed, such as alternative arguments for the same assertion. Counterarguments are not allowed. In sum, the PPD protocol is nondeterministic, multi-move and multi-reply but postponement of replies is not allowed. Dark-side commitments prevent the protocol from having a public semantics.

Most protocol rules refer to the participants' commitments. To start with, challenges, concessions and retractions always concern commitments. Second, a speaker cannot challenge or concede his own commitments, and *question* φ and φ *since* S may not be used if the listener is committed to φ. Furthermore, if a participant has inconsistent commitments, the other participant can demand resolution of the inconsistency by using the *resolve* speech act. Also, if a participant's commitments logically imply an assertion of the other participant but do not contain that assertion, then the initial participant must either concede the assertion or retract one of the implying commitments. Retractions must be successful in that the retracted proposition is not still implied by the speaker's commitments. Finally, the commitments determine the outcome of a dialogue: dialogues terminate after a predetermined number of turns, and the outcome of terminated dialogues is defined as for pure persuasion.

Table 1.2 contains an example dialogue.[3] The first column numbers the turns, and the second contains the moves made in each turn. The other columns contain the assertions and concessions of W and B: the first row contains the initial commitments and the other rows indicate changes in these sets: $+\varphi$ means that φ is added and $-\varphi$ that it is deleted. If the dialogue terminates here, there is no winner, since neither player has conceded any of the other player's assertions or retracted any of his own.

Several points are worth noting about this example. Firstly, B in his first turn moves a complex argument, where the second argument supports a premise of the first: for this reason i is not added to B's assertions. Next, in his second turn, W first concedes j and then asserts j as a premise of an argument; only after the second move has W incurred a burden to defend j if challenged. However, B in his second turn cannot challenge j since B is itself committed to j: if B wants to challenge j, he must first retract j. Note further that after B concedes $f \wedge j \rightarrow a$ in his second turn, his commitments logically imply a, which is an assertion of W. Therefore B must in the same turn either concede a or retract one of the implying commitments. B opts for the latter, retracting f. Next consider B's second move of his second turn: remarkably, B becomes committed to a tautology but W still has the right to challenge it at his third turn. Finally, the example illustrates that the protocol only partly enforces relevance of moves. For instance, at any point a participant could have moved *question* φ for any φ not in the commitments of the listener.

Paul and Olga (ct'd): Let us finally reconstruct our running example in PPD. To start with, Paul's initial claim must now be modelled as an initial assertion in the

[3] In this section the dialogue participants will be denoted with W and B, except if they have proponent/opponent roles throughout the dialogues, in which case they are called P and O.

Table 1.2 An example PPD dialogue

Turn	Moves	A_W	C_W	A_B	C_B
		$\{a\}$	$\{b,c\}$	$\{d,e\}$	$\{f,g\}$
W_1	*why* d *concede* e		 $+e$		
B_1	*why* a d *since* h,i, i *since* j,k			 $+h, h \wedge i \rightarrow d$, $j,k, j \wedge k \rightarrow i$	
W_2	*concede* j *concede* k *why* $j \wedge k \rightarrow i$ *concede* $h \wedge i \rightarrow d$ *why* h a *since* f, j	 $+f, j, f \wedge j \rightarrow a$	$+j$ $+k$ $+h \wedge i \rightarrow d$		
B_2	h *since* $l, l \rightarrow h$ $j \wedge k \rightarrow i$ *since* m *concede* $f \wedge j \rightarrow a$ *retract*$_C$ f			$+l, l \rightarrow k$, $l \wedge (l \rightarrow k) \rightarrow k$ $+m, m \rightarrow (j \wedge k \rightarrow i)$	 $+f \wedge j \rightarrow a$ $-f$

preparatory phase. Since arguments can be incomplete, they can be modelled as in the example's original version. Two features of PPD make a straightforward modelling of the example impossible. The first is that PPD requires that every claim or argument is replied to in the next turn and the second is that explicit counterarguments are not allowed. To deal with the latter, it must be assumed that Olga has also declared an initial assertion, viz. that Paul's car is not safe.

(P_0: *claim* safe — O_0: *claim* $\neg$ safe)
O_1: *why* safe
P_2: safe *since* airbag
P_3: *why* $\neg$ safe — O_4: *concede* airbag
O_5: $\neg$ safe *since* newspaper

Here a problem arises, since Olga now has to either concede or challenge Paul's hidden premise airbag $\rightarrow$ safe. If Olga concedes it, she is forced to also concede Paul's initial claim, since it is now implied by Olga's commitments. If, on the other hand, Olga challenges the hidden premise, then at his next turn Paul must provide an argument for it, which he does not do in our original example. Similar problems arise with the rest of the example. Let us now, to proceed with the example, ignore this 'completeness' requirement of turns.

P_6: *concede* newspaper

Here another problem arises, since PPD does not allow Paul to move his undercutting counterargument against O_5. The only way to attack O_5 is by challenging its unstated premise (newspaper $\rightarrow \neg$ safe).

In sum, two features of PPD prevent a fully natural modelling of our example: the monotonic nature of the underlying logic and the requirement to reply to each claim or argument of the other participant.

1.5.2 Parsons, Wooldridge & Amgoud (2003)

In a series of papers Parsons, Wooldridge & Amgoud have developed an approach to specifying dialogue systems for various types of dialogues. Here the persuasion system of (8) will be discussed.

The system is for dialogues on a single topic between two players called White (W) and Black (B). Dialogues have no context but each participant has a, possibly inconsistent, belief base Σ. The communication language consists of claims, challenges, and concessions; it has no explicit reply structure but the protocol largely conforms to Table 1.1. Claims can concern both individual propositions and sets of propositions. The logical language is propositional. Its logic is an argument-based nonmonotonic logic in which arguments are classical proofs from consistent premises and counterarguments negate a premise of their target. Conflict relations between arguments are resolved with a preference relation on the premises such that arguments are as good as their least preferred premises. Argument acceptability is defined with grounded semantics. In dialogues, arguments cannot be moved as such but only implicitly as *claim* S replies to challenges of another claim φ, such that S is consistent and $S \vdash \varphi$. Finally, the commitment rules are standard and commitments are only used to enlarge the player's belief base with the other player's commitments; they do not constrain move legality nor determine the dialogue's outcome.

An important feature of the system is that the players are assumed to adopt an assertion and an acceptance attitude, which they must respect throughout the dialogue. The attitudes are defined relative to their internal belief base (which remains constant throughout a dialogue) plus both players' commitment sets (which may vary during a dialogue). The following assertion attitudes are distinguished: a *confident* agent can assert any proposition for which he can construct an argument, a *careful* agent can do so only if he can construct such an argument and cannot construct a stronger counterargument, and a *thoughtful* agent can do so only if he can construct an acceptable argument for the proposition. The corresponding acceptance attitudes also exist: a *credulous* agent accepts a proposition if he can construct an argument for it, a *cautious* agent does so only if in addition he cannot construct a stronger counterargument and a *skeptical* agent does so only if he can construct an acceptable argument for the proposition.

It can be debated whether such the requirement to respect these attitudes must be part of a protocol or of a participant's heuristics. According to one approach, a dialogue protocol should only enforce coherence of dialogues (14; 10); according to another approach, it should also enforce rationality and trustworthiness of the agents engaged in a dialogue (8). The second approach allows protocol rules to refer to an agent's internal belief base and therefore such protocols do not have a

public semantics. The first approach does not allow such protocol rules and instead regards assertion and acceptance attitudes as an aspect of agent design.

The formal definition of the persuasion protocol is as follows.

Definition 7. (PWA persuasion protocol) A move is legal iff it does not repeat a move of the same player, and satisfies the following procedure:

1. W claims φ.
2. B concedes φ if its acceptance attitude allows, if not B asserts $\neg\varphi$ if its assertion attitude allows it, or otherwise challenges φ.
3. If B claims $\neg\varphi$, then goto 2 with the roles of the players reversed and $\neg\varphi$ in place of φ.
4. If B has challenged, then:
   a. W claims S, an argument for φ;
   b. Goto 2 for each $s \in S$ in turn.
5. B concedes φ if its acceptance attitude allows, or the dialogue terminates.

Dialogues *terminate* as specified in condition 5, or when the move required by the procedure cannot be made, or when the player-to-move has conceded all claims made by the hearer.

No win and loss functions are defined, but the possible outcomes are defined in terms of the propositions claimed by one player and conceded by the other.

This protocol is unique-move except that if one element of a *claim S* move is conceded, another element may be replied-to next. Also, it is unique-reply except that each element of a *claim S* move can be separately challenged or conceded. The protocol is deterministic in $\mathscr{L}_c$ but not fully deterministic, since if a player can construct more than one argument for a challenged claim, he has a choice.

Let us first consider some simple dialogues that fit this protocol.

Example 1. First, let $\Sigma_W = \{p\}$ and $\Sigma_B = \varnothing$. Then the only legal dialogue is:

W_1: *claim* p, $\quad B_1$: *concede* p

B_1 is B's only legal move, whatever its acceptance attitude, since after W_1, B must reason from $\Sigma_B \cup C_W(d_1) = \{p\}$ so that B can construct the trivial argument $(\{p\}, p)$. Here the dialogue terminates.

This example illustrates that since the players must reason with the commitments of the other player, they can learn from each other. However, the next example illustrates that the same feature sometimes makes them learn too easily.

Example 2. Assume $\Sigma_W = \{q, q \rightarrow p\}$ and $\Sigma_B = \{\neg q\}$, where all formulas are of the same preference level.

W_1: *claim* p

Now whatever her acceptance attitude, B has to concede p since she can construct the trivial argument $(\{p\}, p)$ for p while she can construct no argument for $\neg p$. Yet B has an attacker for W's only argument for p, namely, $(\{\neg q\}, \neg q)$, which attacks $(\{q, q \rightarrow p\}, p)$ and is not weaker than its target. So even though p is not acceptable on the basis of the agents' joint knowledge, W_1 can win a dialogue about p.

This example thus illustrates that if the players must reason with the other player's commitments, one player can sometimes 'force' an opinion onto the other player by simply making a claim. A simple solution to this problem is to restrict the information with which agent reason to their own beliefs and commitments. A more refined option is to assume that the agents have knowledge about the reliability of information sources and to let them use it in the acceptance policies.

Paul and Olga (ct'd): Finally, our running example can be modelled in this approach as follows. Let us give Paul and Olga the following beliefs:

$\Sigma_W = \{\text{airbag}, \text{airbag} \rightarrow \text{safe}, \neg(\text{newspaper} \rightarrow \neg\ \text{safe})\}$
$\Sigma_B = \{\text{newspaper}, \text{newspaper} \rightarrow \neg\ \text{safe}\}$

(Note that Paul's undercutter must now be formalised as the negation of Olga's material implication.) Assume that all these propositions are equally preferred. We must also make some assumptions on the players' assertion and acceptance attitudes. Let us first assume that Paul is thoughtful and skeptical while Olga is careful and cautious, and that they only reason with their own beliefs and commitments.

P_1: *claim* safe O_2: *claim* $\neg$ safe

Olga could not challenge Paul's main claim as in the example's original version, since she can construct an argument for '$\neg safe$', while she cannot construct an argument for 'safe'. So she had to make a counterclaim. Now since players may not repeat moves, Paul cannot make the remove required by the protocol and his assertion attitude, namely, claiming 'safe', so the dialogue terminates without agreement.

Let us now assume that the players must also reason with each others commitments. Then the dialogue evolves as follows:

P_1: *claim* safe O_2: *concede* safe

Olga has to concede, since she can use Paul's commitment to construct the trivial argument $(\{\text{safe}\}, \text{safe})$, while her own argument for '$\neg$ safe' is not stronger. So here the dialogue terminates with agreement on 'safe', even though this proposition is not acceptable on the basis of the players' joint beliefs.

So far, neither of the players could develop their arguments. To change this, assume now that Olga is also thoughtful and skeptical, and that the players reason with each others commitments. Then:

P_1: *claim* safe O_2: *why* safe

Olga could not concede, nor could she state her argument for $\neg$ safe since it is not preferred over its attacker $(\{\text{safe}\},\text{safe})$. So she had to challenge.

P_3: *claim* $\{\text{airbag}, \text{airbag} \rightarrow \text{safe}\}$

Now Olga can create a (trivial) argument for 'airbag' by using Paul's commitments, but she can also create an argument for its negation by using her own beliefs. Neither is acceptable, so she must challenge. Likewise for the second premise, so:

	O_4: *why* airbag
P_5: *claim* $\{\text{airbag}\}$	O_6: *why* airbag $\rightarrow$ safe
P_7: *claim* $\{\text{airbag} \rightarrow \text{safe}\}$	

Here the nonrepetition rule makes the dialogue terminate without agreement. Note that only Paul could develop his arguments. To give Olga a chance to develop her arguments, let us make her careful and skeptical while the players still reason with each others commitments. Then:

P_1: *claim* safe O_2: *claim* $\neg$ safe

In the new dialogue state Paul's argument for 'safe' is not acceptable any more, since it is not preferred over its attacker ($\{\neg$ safe$\}$, $\neg$ safe). So he must challenge.

P_3: *why* $\neg$ safe O_4: *claim* $\{$newspaper, newspaper $\rightarrow \neg$ safe $\}$

Although Paul can construct an argument for Olga's first premise, namely, ($\{\neg$(newspaper $\rightarrow \neg$ safe'$\}$, safe), it is not acceptable since it is not preferred over its attacker based on Olga's second premise. So he must challenge.

P_5: *why* newspaper O_6: *claim* $\{$newspaper$\}$

Olga had to reply with a (trivial) argument for her first premise, after which Paul cannot repeat his challenge, so he has to go to the second premise of O_4. Based on his beliefs and Olga's commitments he can construct (trivial) arguments both for and against it and neither of these is acceptable. So he must again challenge.

P_7: *why* newspaper $\rightarrow \neg$ safe O_8: *claim* $\{$newspaper $\rightarrow \neg$ safe$\}$

Here the nonrepetition rule again makes the dialogue terminate without agreement. In this dialogue only Olga could develop her arguments (although she could not state her second counterargument).

In conclusion, the PWA persuasion protocol leaves little room for choice and exploring alternatives, and induces one-sided dialogues in that at most one side can develop their arguments for a certain issue. Also, the examples suggest that if a claim is accepted, it is accepted in the first 'round' of moves (but this should be formally verified). On the other hand, the strictness of the protocol induces short dialogues which are guaranteed to terminate, which promotes efficiency. Also, thanks to the strong assumptions on the logic and the participants' beliefs and reasoning behaviour, PWA have been able to prove several interesting properties of their protocols. Finally, without the requirement to respect the assertion and acceptance attitudes the protocol would be much more liberal while still enforcing some coherence.

1.5.3 Prakken (2005)

In (10) I proposed a formal framework for systems for two-party persuasion dialogues and instantiated it with some example protocols. The participants have proponent and opponent role, and their beliefs are irrelevant to the protocols, so that these have a public semantics. Dialogues have no context. The framework abstracts from the communication language except for an explicit reply structure. It also abstracts from the logical language and the logic, except that the logic is assumed to

be argument-based and to conform to grounded semantics and that arguments are trees of deductive and/or defeasible inferences, as in e.g. (9).

A main motivation of the framework is to ensure focus of dialogues while yet allowing for freedom to move alternative replies and to postpone replies. This is achieved with two main features of the framework. Firstly, $\mathscr{L}_c$ has an explicit reply structure, where each move either *attacks* or *surrenders to* its target. An example $\mathscr{L}_c$ of this format is displayed in Table 1.3. Secondly, winning is defined for each dialogue, whether terminated or not, and it is defined in terms of a notion of *dialogical status* of moves. The *dialogical status* of a move is recursively defined as follows, exploiting the tree structure of dialogues generated by the reply structure on $\mathscr{L}_c$. A move is *in* if it is surrendered or else if all its attacking replies are *out*. (This implies that a move without replies is *in*). And a move is *out* if it has a reply that is *in*. Then a dialogue is (currently) won by the proponent if its initial move is *in* while it is (currently) won by the opponent otherwise.

Table 1.3 An example L_c in Prakken's framework

Acts	Attacks	Surrenders
claim φ	*why* φ	*concede* φ
φ *since* S	*why* $\psi (\psi \in S)$	*concede* ψ $(\psi \in S)$
	φ' *since* S' (φ' *since* S' defeats φ *since* S)	*concede* φ
why φ	φ *since* S	*retract* φ
concede φ		
retract φ		

Together, these two features of the framework support a notion of relevance that ensures focus while yet leaving a degree of freedom: a move is *relevant* just in case making its target *out* would make the speaker the current winner. Termination is defined as the situation that a player is to move but has no legal moves. Various results are proven about the relation between being the current winner of a dialogue and what is defeasibly implied by the arguments exchanged during the dialogue.

As for dialogue structure, the framework allows for all kinds of protocols. The instantiations of (10) are all multi-move and multi-reply; one of them has the communication language of Table 1.3 and is constrained by the requirement that each move be relevant. This makes the protocol immediate-response, which implies that each turn consists of zero or more surrenders followed by one attacker. Within these limits postponement of replies is allowed, sometimes even indefinitely.

Let us next discuss some examples, assuming that the protocol is further instantiated with Prakken & Sartor's argument-based version of prioritised extended logic programming (12). This logic uses grounded semantics and supports arguments about rule priorities. (The examples below should speak for themselves so no formal definitions about the logic will be given. Note that since the rules are logic-programming rules, they do not satisfy contraposition or modus tollens. Rule

connectives are tagged with a rule name, which is needed to express rule priorities in the object language). Consider two agents with the following belief bases:

$$\Sigma_P = \{p, p \Rightarrow_{r_1} q,\ q \Rightarrow_{r_2} r,\ p \wedge s \Rightarrow_{r_3} r_2 > r_4\}$$
$$\Sigma_O = \{t, t \Rightarrow_{r_4} \neg r\}.$$

Then the following is legal in (10)'s so-called relevant protocol (with each move its target is indicated between square brackets):

$P_1[-]$: *claim* r	$O_2[P_1]$: *why* r
$P_3[O_2]$: r *since* $q, q \Rightarrow r$	$O_4[P_3]$: *why* q
$P_5[O_4]$: q *since* $p, p \Rightarrow q$	$O_6[P_5]$: *concede* $p \Rightarrow q$
	$O_7[P_5]$: *why* p

(Note that unlike in (8) but like in (14), arguments can be stepwise built in several moves.) Here P has several allowed moves, viz. retracting any of his argument premises or his claim, or giving an argument for p. All these moves are relevant but if P makes any retraction then an argument for p ceases to be relevant, since it cannot make P the current winner. Moreover, if P retracts r as a reply to P_1 then the dialogue terminates with a win for O.

O could at all points after P_3 have moved her argument against r. For instance:

	$O_7[P_3]$: $\neg r$ *since* $t, t \Rightarrow \neg r$
$P_8[O_7]$: $r_2 > r_4$ *since* $p, s, p \wedge s \Rightarrow r_1 > r_4$	

P_8 is a priority argument which makes P_3 strictly defeat O_7 (note that the fact that s is not in P's own belief base does not make the move illegal). At this point, P_1 is *in*; O has various allowed moves, viz. challenging or conceding any (further) premise of P's arguments, moving a counterargument to P_5 or a second counterargument to P_3, and conceding P's initial claim.

This example shows that the participants have much more freedom in this system than in the one of (8), since they are not bound by assertion and acceptance attitudes and the protocol is structurally less strict. The downside of this is that dialogues can be much longer, that the participants can lie and that they can prevent losing by simply continuing to attack the other participant.

Another drawback of the present approach is that not all natural-language dialogues have an explicit reply structure. For example, often one player tries to extract seemingly irrelevant concessions from the other player with the aim to lure her into a contradiction, as in as in the following witness cross-examination dialogue:

Witness: Suspect was at home with me that day.
Prosecutor: Are you a student?
Witness: Yes.
Prosecutor: Was that day during summer holiday?
Witness: Yes.
Prosecutor: Aren't all students away during summer holiday?

In (14) such dialogues can be modelled with the *question* locution but at the price of decreased coherence and focus.

Paul and Olga (ct'd): Let us finally model our running example in this protocol. Figure 1.2 displays the dialogue tree, where moves within solid boxes are *in* and moves within dotted boxes are *out*.

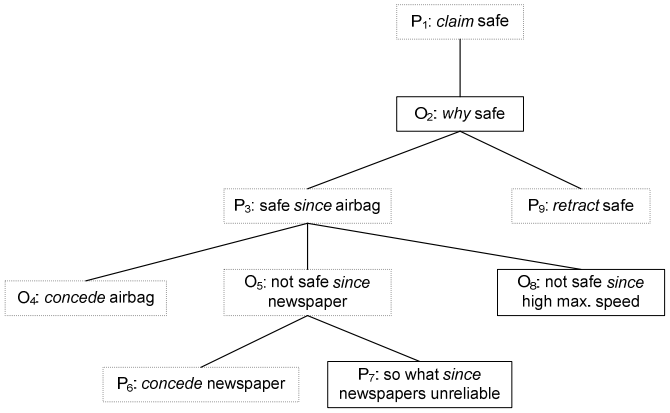

Fig. 1.2 The example dialogue in Prakken's approach

As can be easily checked, this formalisation captures all aspects of the example's original version, except that arguments have to be complete and that counterarguments cannot be introduced by a counterclaim. (But other instantiations of the framework may be possible without these limitations.)

1.6 Conclusion

In this chapter a formal framework for dialogue systems for persuasion was proposed, which was then used to critically discuss three systems from the literature. Concluding, we can say that the formal study of persuasion dialogue has resulted in a number of interesting dialogue systems, some of which have been applied in insightful case studies or applications. On the other hand, there is still much room for refining or extending the various sytems with, for example, more refined communication languages or with different modes of reasoning, such as probabilistic, case-based or coherence-based reasoning. Also, the integration of persuasion with other types of dialogues should be studied. Another important research issue is the study of strategies and heuristics for individual participants and how these interact with the protocols to yield certain properties of dialogues. One aspect of such studies is the development of quality measures for dialogues as to how well they satisfy certain desirable properties. More generally, a formal metatheory of systems, their interrelations and their combinations with agent models is still in its early stages.

Perhaps the main challenge in tackling all these issues is how to reconcile the need for flexibility and expressiveness with the aim to enforce coherent dialogues. The answer to this challenge may well vary with the nature of the context and application domain, and a precise description of the grounds for such variations would provide important insights in how dialogue systems for persuasion can be applied.

References

[1] J. Barwise and L. Moss. *Vicious Circles*. Number 60 in CSLI Lecture Notes. CSLI Publications, Stanford, CA, 1996.

[2] G. Brewka. Dynamic argument systems: a formal model of argumentation processes based on situation calculus. *Journal of Logic and Computation*, 11:257–282, 2001.

[3] L. Carlson. *Dialogue Games: an Approach to Discourse Analysis*. Reidel Publishing Company, Dordrecht, 1983.

[4] T. Gordon. The Pleadings Game: an exercise in computational dialectics. *Artificial Intelligence and Law*, 2:239–292, 1994.

[5] C. Hamblin. *Fallacies*. Methuen, London, 1970.

[6] R. Loui. Process and policy: resource-bounded non-demonstrative reasoning. *Computational Intelligence*, 14:1–38, 1998.

[7] J. Mackenzie. Question-begging in non-cumulative systems. *Journal of Philosophical Logic*, 8:117–133, 1979.

[8] S. Parsons, M. Wooldridge, and L. Amgoud. Properties and complexity of some formal inter-agent dialogues. *Journal of Logic and Computation*, 13, 2003. 347-376.

[9] J. Pollock. *Cognitive Carpentry. A Blueprint for How to Build a Person*. MIT Press, Cambridge, MA, 1995.

[10] H. Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation*, 15:1009–1040, 2005.

[11] H. Prakken. Formal systems for persuasion dialogue. *The Knowledge Engineering Review*, 21:163–188, 2006.

[12] H. Prakken and G. Sartor. Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-classical Logics*, 7:25–75, 1997.

[13] D. Walton. *Logical dialogue-games and fallacies*. University Press of America, Inc., Lanham, MD., 1984.

[14] D. Walton and E. Krabbe. *Commitment in Dialogue. Basic Concepts of Interpersonal Reasoning*. State University of New York Press, Albany, NY, 1995.

[15] J. Woods and D. Walton. Arresting circles in formal dialogues. *Journal of Philosophical Logic*, 7:73–90, 1978.

[16] T. Yuan, D. Moore, and A. Grierson. A human-computer dialogue system for educational debate: A computational dialectics approach. *International Journal of Artificial Intelligence in Education*, 18:3–26, 2008.