JAN VAN KUPPEVELT

INFERRING FROM TOPICS

Scalar Implicatures as Topic-Dependent Inferences

1. Introduction[1]

This paper focuses on the specific inferential phenomena, known as generalized conversational implicatures, which arise from Grice's first submaxim of Quantity (Grice 1967/75). Particular attention is paid to the subclass of so-called *scalar implicatures* (Gazdar 1979; Horn 1972). This subclass has been widely discussed and is recognized as the type of pragmatic inference that has probably been most systematically accounted for. However, as is generally acknowledged, inferences of this type are not without problems, especially how they are actually generated. These problems are mainly due to the difficulty of defining adequately the notion of linguistic scales underlying their explanation.

Central to scale definition are the points of *scale activation*, *scale reduction*, and, in particular, *scale ordering* and *scale coherence*. As indicated by Harnish (1979), Hirschberg (1985) and others, the activation of scales must be restricted in order to avoid an overgeneration of inferences based on them. The second point – scale reduction – is probably the most difficult one (see, in particular, Rooth 1992). Scale reduction considers the contextual constraints on the (original) number of elements defining a scale. The much discussed point of scale ordering (see Fauconnier 1975a/b; Gazdar 1979; Hirschberg 1985; Horn 1972/89) refers to a general ordering criterion for the elements of a scale. This may be a semantic, pragmatic or other type of ordering. The last point – scale coherence – is fairly diverse (see Hirschberg 1985; Horn 1972/89; Levinson 1983). Problems observed in this respect include the *hierarchy problem*, which implies that in specific cases the elements on a scale cannot be hierarchically ordered without leading to incorrect scalar predictions, and the problems of *scale overlap*, *scale direction* and *scale partitioning*.

Directly related to all these problems are the point of the selection of a value on a scale and the context-dependency of a satisfactory selection. Both points relate to an adequate prediction of a scalar inference. Largely disregarded in the literature is the fact that the selection of a scale value

---

[1] The research reported here was supported by the Royal Netherlands Academy of Arts and Sciences (KNAW). I also wish to thank the two anonymous referees of *Linguistics and Philosophy* for their useful comments on an earlier version of this paper.

Linguistics and Philosophy **19**: 393–443, 1996.

may not only consist of a unique value but also of a less specific, non-unique value, comprising a number of scale values. Generally, in the latter case, no inference is generated. But the context preceding an inference-inducing sentence may also be such that it weakens the requirements for a satisfactory selection, thereby implying that no inference is induced in such cases either.

In this paper we propose a definition of linguistic scales in terms of a uniform topic notion comprising both the notion of sentence topics and that of topics of larger discourse units. The aim of this definition is to provide a solution to the problems cited above. The argument presented favors linguistic scales as ordered topic ranges introduced by higher-order or lower-order topic-forming questions, either explicit or implicit. This implies that the notion of linguistic scales is directly related to that of the segmentation structure of discourse which is considered to be in correspondence with the discourse-internal topic-comment structure. Linguistic scales thus give rise to inferences generated on different discourse levels. The inferences based on them are not only associated with single utterances, as is generally assumed, but are also generated as the result of larger discourse units, including those comprising the discourse as a whole.

The theoretical framework that underlies our proposal assumes a direct relationship between topics, questions and discourse structure. In Van Kuppevelt (1991/95a) it is hypothesized that a discourse derives its structural coherence from the internal topic-comment structure which results from the process of the contextual induction of explicit and implicit topic-forming questions. On the global level a linguistic scale is defined by the higher-order topic-forming question answered by a larger discourse unit. The inference generated on the basis of this scale is the one related to the whole discourse segment. On the local level, on the other hand, different linguistic scales and the inferences based on them can be associated with individual sentences that answer a (sub)topic-forming question. We demonstrate the way in which the inferential output of scalar inferences generated on the local level amounts to the scalar inference on the corresponding global level. We present two generalizations for computing such higher-order scalar inferences: one concerning the *actual inducing context* which does not necessarily coincide with the discourse unit as a whole, and the other, the *moment of induction*, which does not necessarily coincide with the moment at which the implicature inducing context is generated.

As to scale activation, we put forward the view that the contextual

restrictions which hold for implicature generation are determined by the explicit or implicit topic-forming question answered by the implicature inducing discourse unit. With reference to this, we argue that contextually determined inferences of this type have in fact an inferential status different to the weaker pragmatic status that is generally assumed. In this respect, we discuss Horn's (1972) distinction between two kinds of elimination of a scalar implicature, and we give an adequate response to Gazdar's (1979) arguments that scalar implicatures cannot be (semantic) entailments.

The definition of linguistic scales as ordered topic ranges directly presents a solution to the remaining scale problems. Scale reduction is analyzed as a function of the completion task of subquestions. This task consists of a further reduction of the original topic range defined by the higher-order topic-forming question. As far as an ordering criterion for scales is concerned, we propose a general criterion in terms of answer informativeness. Finally, the problem of scale coherence, in particular the question of scale overlap, scale direction and scale partitioning, is accounted for in terms of whether the scale values in question share the same topic.

The above mentioned problems directly related to scale definition are also accounted for in terms of topic-forming questions. The selection of unique or non-unique scale values is related to that of satisfactory or unsatisfactory answers to topic-forming questions. It is shown that the satisfactoriness of an answer depends, among other things, on the superordinated higher-order question. As is made clear, both points are central to our claim that cardinals do not, as is generally assumed, differ principally from non-cardinals with respect to the generation of scalar inferences.

We will start the analysis with a brief outline of the framework that presents an account of hierarchical discourse structure in terms of topic-forming questions (Section 2). A characterization will be given of the question-based topic notion, the notion of topic range that is implied by this and the reduction of such a range to a unique value as the result of the process of explicit and implicit questioning in discourse. We continue with the main subject of a topical account of scalar inferences in terms of topic-forming questions (Section 3). We first account for the view that scalar implicatures are in fact determined by topic-forming questions. Thereafter, we explain the notion of linguistic scales as ordered topic ranges. Finally, an account is given of the generation of scalar inferences on different discourse levels.

2. Topicality and Questioning

2.1. *The Question-Based Topic Notion*

The process of questioning in discourse which gives rise to scalar inferences presupposes a non-trivial relation between topic hierarchy and (hierarchical) discourse structure. Central to this is the view that the organization of discourse segments is in agreement with the discourse-internal topic-comment structure which results from this questioning process. As is argued in Van Kuppevelt (1991 and elsewhere), this implies that the topic of a discourse unit is determined by the explicit or implicit question it answers and that structural relations in discourse are defined by the relations between these topic-providing questions. The framework presupposes a context-dependent and question-based topic notion which accounts for sentence topics and discourse topics in a uniform way.[2]

By definition, a topic $\mathbf{T_p}$ is *that which is being questioned* by means of a contextually induced explicit or implicit question $\mathbf{Q_p}$. The corresponding comment $\mathbf{C_p}$ is provided by answer $\mathbf{A_p}$. $\mathbf{C_p}$ is *that which is asked for* by $\mathbf{Q_p}$. If (the speaker assumes) $\mathbf{A_p}$ is sufficiently satisfying to that which is asked for, $\mathbf{T_p}$ is closed off. If not, as will be explained later, $\mathbf{A_p}$ will give rise to subquestioning.[3,4]

Topic $\mathbf{T_p}$, that which is being questioned, is semantically characterized as the *intension* of the topic term in the syntactic analysis of the question, e.g. the intension of *the one who is laughing* functioning as the topic term in the syntactic analysis of the question given in (1).

(1)a. Q_1 Who is (the one who is) laughing?
$T_1(S_{act}) = -?$

b. A_1 Alan is laughing.
$T_1(S_{act}) = C_1$

In an ongoing discourse this topic term denotes a textually given or evoked discourse address for which no unique extensional counterpart exists in

---

[2] For other question-based topic notions see, e.g., Bartsch (1976), Klein and Von Stutterheim (1987), Stout (1986), and Vennemann (1975).

[3] In the context of addressee-oriented discourse, implicit questions are defined as those questions the speaker anticipates the addressee asking as the result of the preceding context.

[4] Elsewhere we present an outline of an algorithm for implicit question reconstruction (Van Kuppevelt 1991). Central to it is the reconstruction of implicit questions on the basis of certain formal characteristics of the (spoken) text, including accent distribution, specific syntactic structures like cleft and pseudo-cleft structurcs, and word order. However, a fully adequate reconstruction method also requires that other, not strictly (con)textual factors are taken into account, especially those related to the interaction of given contextual information and assumed background and situational knowledge.

the contextual domain. Topic T_1 is the set of possible extensions of this term. Assuming that someone is laughing and assuming a contextual domain D = {Alan, Brian} this topic is defined as follows:

(2) $T_1 = \{\langle S_1, \{\text{Alan}\}\rangle, \langle S_2, \{\text{Brian}\}\rangle, \langle S_3, \{\text{Alan, Brian}\}\rangle\}$

A comment $\mathbf{C_p}$, that which is asked for by the corresponding question $\mathbf{Q_p}$, is the *extension* $\mathbf{T_p(S_{act})}$ of the topic term in the actual situation, e.g. the extension {Alan} provided by answer A_1 in (1b). This extensional value is selected by A_1 from the topic set T_1. Question-answer pairs can thus be represented extensionally as in (1) in which the full answer A_1 provides the actual extension of the topic term, namely $\langle S_{act}, T_1(S_{act})\rangle$.[5]

2.2. *Topic Ranges and Unique Determination*

A central issue of the theory outlined here is that topic-forming questions are induced as the result of textually given or evoked *indeterminacies* or so-called *question locations*. If it is textually given, a question location is a non-uniquely referring term which, because of its referential ambiguity, is made the subject of questioning and, as a consequence of this, becomes a topic expression. At the moment of questioning the extension of this term, i.e. $\mathbf{T_p(S_{act})}$, is (still) un(der)determined, implying that the actual topic range $\boldsymbol{\rho}'\mathbf{(T_p)}$ of which $\mathbf{T_p(S_{act})}$ is an element does not yet contain a unique value.

(3) $T_p(S_{act}) \in \rho'(T_p) \qquad |\rho'(T_p)| > 1$

The actual topic range $\boldsymbol{\rho}'\mathbf{(T_p)}$ comprises the original or remaining set of possible extensional values $\mathbf{T_p(S_i)}$ for the topic term in question. In case of example (2) $\rho'(T_1)$ is, at the moment of questioning, identical to the original topic range: $\rho'(T_1)$ = {{Alan}, {Brian}, {Alan, Brian}}.

A reduction of the un(der)determinedness of the actual topic extension $\mathbf{T_p(S_{act})}$ is realized by an answer to the corresponding question $\mathbf{Q_p}$. If satisfactory (and if no disturbance occurs in the questioning process), it involves the unique determination of $\mathbf{T_p(S_{act})}$.[6] In that case the actual

---

[5] Instead of a propositional account of questions and answers (e.g., Belnap 1982; Groenendijk and Stokhof 1984; Hamblin 1973; Karttunen 1977), we assume an individualistic one (e.g., Hausser 1983; Scha 1983; Tichý 1978). The analysis is in agreement with the view explicit in, e.g., Belnap and Steel (1976) that the topic ('subject') of a question is a set of alternatives.

[6] Other configurations of unique determination include those in which a satisfactory answer is *inferred* from an apparently unsatisfactory one and those obtained by *topic narrowing* and *topic weakening* processes (Van Kuppevelt 1994).

topic range is reduced to only one value: $|\rho'(T_p)| = 1$. In the case of (1b), for example, the answer results in the unique determination $T_1(S_{act}) \in \{\{Alan\}\}$ (or: $T_1(S_{act}) = \{Alan\}$) of the actual topic extension.

The unique determination of the actual topic extension $T_p(S_{act})$ implies that the necessary condition for topichood, namely the un(der)determinedness of a textually given or evoked question location, is no longer being met. This automatically results in topic closure. However, topic closure may also be forced as the result of an epistemic limitation on the part of the answerer, implying that he does not *know* a satisfactory, uniquely determining answer to the question. Such a situation is frequently marked by a phrase such as 'All I know is that . . .'.

As will be explained in the next section, the state of unique determination is usually not reached in one step but involves all, or a considerable part of, the discourse. This is the case in hierarchically structured discourses underlying complex topic processes which necessarily involve answers to subtopic-forming subquestions.

2.3. *Topic Hierarchy and Discourse Structure*

2.3.1. *Main Topic-Constituting Questions*

The questioning process underlying the discourse production process controls the development of the discourse and provides it with its mostly hierarchical segmentation structure and corresponding topic-comment structure. In this respect it is demonstrated that the main structure of a coherent discourse results from the contextual induction of two functionally different types of topic-forming questions, i.e. *main, topic--constituting questions* and *subtopic-constituting subquestions*.[7]

Every main, explicit or implicit, topic-constituting question Q_p is induced as the result of a linguistic or non-linguistic *feeder* F_i. The function of a linguistic feeder F_i is to initiate or re-initiate the process of questioning in discourse. It may be a relatively large discourse unit or just a single sentence, e.g. the opening sentence of a discourse or a sentence which serves to continue the conversation when no more questions are induced by the preceding discourse. Together with associated background knowledge a feeder F_i gives rise to a *discourse topic* DT_i which, by definition, consists of all the main, higher-order topics constituted as the result of F_i.

Consider the following example in which a feeder gives rise to a

[7] The distinction between main structure and side structure is accounted for in Van Kuppevelt (1995b).

relatively simple, non-hierarchical discourse resulting from just one main, explicit topic-constituting question.

(4)a. F_1 A: A well-known subsidy book publisher is searching for manuscripts.
Q_1 B: What kind of manuscripts?
A_1 A: Fiction and non-fiction.

b. Q_1: What kind of manuscripts?

$$T_1(S_{act}) \in \{X \mid X \subseteq \text{TYPES\_OF\_MANUSCRIPTS}\}$$

A_1: Fiction and non-fiction.

$$T_1(S_{act}) = \{\text{Fiction, Non-fiction}\}!^8$$

Question Q_1 introduces an undetermined actual topic extension $T_1(S_{act})$. At the moment of questioning $T_1(S_{act})$ is identical to one of the many possible combinations of types of manuscripts, as indicated by the topic range $\{X \mid X \subseteq \text{TYPES\_OF\_MANUSCRIPTS}\}$. The unique determination of $T_1(S_{act})$ is achieved here in a single step by means of answer A_1. It is assumed that A_1 is a satisfactory answer because there is no epistemic limitation and no additional questions arise as the result of this answer.

2.3.2. *Subtopic-Constituting Subquestions*

If the answer to an explicit or implicit topic-constituting question is (assumed to be) unsatisfactory for the addressee, a (recursive) process of subquestioning involving the constitution of subordinate topics will be initiated. Processes of subquestioning are induced if the unsatisfactory answer has not resulted in a full reduction of the topic range implying that the information state consisting of the unique determination of the actual topic extension has not yet been reached.

Subtopic-constituting subquestions are contextually induced, in a recursive way, as the result of unsatisfactory answers with the purpose of completing them to satisfactory ones. This completion function consists in achieving a further reduction of the underdeterminedness of the actual extension of the main topic (expression).

An explicit or implicit subquestion $\mathbf{Q_p}$ is contextually induced as the result of a *quantitatively* or *qualitatively* unsatisfactory answer $\mathbf{A_{p-n}}$ given to a preceding higher-order question $\mathbf{Q_{p-n}}$. In the former case the unsatisfactory answer results in a quantitative underdeterminedness of the actual

---

[8] In the analyses given topic closure is indicated by an exclamation mark.

topic extension $\mathbf{T_{p\_n}(S_{act})}$, implying that the comment value provided by $\mathbf{A_{p\_n}}$ is incomplete. In the latter case the underdeterminedness of $\mathbf{T_{p\_n}(S_{act})}$ is of a qualitative nature, e.g. because $\mathbf{A_{p\_n}}$ is not specific enough.

Consider first the following variant of example (4).

(4)'a. F_1 A: A well-known subsidy book publisher is searching for manuscripts.
Q_1 B: What kind of manuscripts?
A_1 A: Fiction and non-fiction will be considered.
Q_2 B: What else?
A_2 A: Poetry, juvenile, travel, scientific, specialized and even controversial subjects.

b. Q_1: What kind of manuscripts?

$$T_1(S_{act}) \in \{X \mid X \subseteq \text{TYPES\_OF\_MANUSCRIPTS}\}$$

A_1: Fiction and non-fiction will be considered.

$$T_1(S_{act}) \in \{T_1(S_i) \in \rho'(T_1) \mid \{\text{Fiction Non-fiction}\} \subseteq T_1(S_i)\}^9$$

Q_2: What else?

$$T_2(S_{act}) \in \{X \mid X \subseteq (\text{TYPES\_OF\_MANUSCRIPTS} - \{\text{Fiction, Non-fiction}\}) \wedge |X| \geqslant 1\})$$

A_2: Poetry, juvenile, travel, scientific, specialized and even controversial subjects.

$$T_2(S_{act}) = \{\text{Poetry, Juvenile, Travel, Scientific, Specialized, Controversial subjects}\}!$$

$$T_1(S_{act}) = \{\text{Fiction, Non-fiction, Poetry, Juvenile, Travel, Scientific, Specialized, Controversial subjects}\}!$$

The occurrence of question Q_2 indicates that the inducing unsatisfactory answer A_1 is unsatisfactory in a quantitative way. The determination of the actual topic extension $T_1(S_{act})$ is now realized in two stages. First, the unsatisfactory answer A_1 reduces the original undeterminedness of $T_1(S_{act})$ to a topic range containing only those (possible) extensions which include the incomplete value {Fiction, Non-fiction}. Second, answer A_2 further

---

[9] An unsatisfactory, *incomplete answer* A_r is formally represented as A_r: $T_r(S_{act}) \in \{T_r(S_i) \in \rho'(T_r) \mid C_r' \subseteq T_r(S_i)\}$, whereby C_r' is the comment value as mentioned in answer A_r.

reduces the set of remaining extensions to one unique value, implying the determination of the actual topic extension $T_1(S_{act})$. After an answer has been given to subquestion Q_2, not only topic T_1 but also subtopic T_2 is closed off. The actuality of the former is continued while subquestion Q_2 is being asked.

The following variant of (4) illustrates the contextual induction of a subquestion as the result of a qualitatively unsatisfactory answer.

(4)″a. F_1 A: A well-known subsidy book publisher is searching for potentially successful manuscripts.
Q_1 B: What kind of manuscripts?
A_1 A: Only fiction.
Q_2 B: What kind of fiction?
(I heard that the success of some types of fiction has been decreasing in recent months.)
A_2 A: Both novels and short stories.

b. Q_1: What kind of manuscripts?

$$T_1(S_{act}) \in \{X \mid X \subseteq \text{TYPES\_OF\_MANUSCRIPTS}\}$$

A_1: Only fiction.

$$T_1(S_{act}) \in X \mid X \subseteq \text{FICTION} \wedge |X| \geqslant 1\}^{10}$$

Q_2: What kind of fiction?
(I heard that the success of some types of fiction has been decreasing in recent months.)

$$T_2(S_{act}) \in \{X \mid X \subseteq \{\text{Novels, Short stories, . . .}\} \wedge |X| \geqslant 1\}$$

A_2: Both novels and short stories.

$$T_2(S_{act}) = \{\text{Novels, Short stories}\}!$$

$$T_1(S_{act}) = \{\text{Novels, Short stories}\}!$$

As in the preceding case, the determination of the actual topic extension $T_1(S_{act})$ involves two steps. However, in this case answer A_1 is unsatisfactory because it is not specific enough.[11] The unique determination of $T_1(S_{act})$ is achieved only after subquestion Q_2 has been answered. Both in this and the preceding case the segmentation structure of the discourse

---

[10] An unsatisfactory, *non-specific answer* A_r is formally represented as $A_r: T_r(S_{act}) \in \rho\{T_r(S_i) \in \rho(T_r) \mid T_r(S_i) \subseteq Y \wedge |T_r(S_i)| \geqslant n\}$, whereby Y and n are given by A_r.

[11] In this paper we focus mainly on this type of qualitative underdeterminedness. Another type is discussed briefly in Section 3.3.1 and more extensively in Van Kuppevelt (1994).

results from the process of questioning and is in agreement with the discourse-internal topic-comment structure.

Both types of subquestions are thus subservient to the process of answering a main, topic-constituting question. Inherent to their completion function is the fact that the satisfactoriness of their answers depends on the goal of the main, superordinating question.[12] If this goal is satisfied, the answer to the subquestion is also satisfactory. Often, this implies that the satisfactory answer and, as a consequence, the topic range introduced by the corresponding subquestion are restricted, compared to what they would be in other, non-hierarchical contexts.[13] Main, higher-order questions thus can give rise to so-called processes of topic narrowing and topic weakening, implying a quantitative or qualitative reduction of the topic range associated with the subquestion. Consider in this respect the following variant of Kempson's (1986) example, analyzed here in terms of question-answer structure (angled brackets indicate the implicit character of a question).

(5) F_1 A: I'm a mother.
Q_1 Do I get a fixed amount of state benefit?
A_1 B: If you have at least two children, you get a fixed amount of state benefit.
$\langle Q_2 \rangle$ ⟨How many children do you have?⟩
A_2 A: I have two children. (In fact I have four.)

The example illustrates that a satisfactory answer to the implicit subquestion Q_2 depends on the context, namely the higher-order, topic-constituting question Q_1. Subquestion Q_2 is subservient to Q_1, the latter of which is satisfactorily answered for both A and B if they know if A gets a fixed amount of state benefit. This is already the case if they know that A has *at least* two children. An 'exactly two' interpretation of the cardinal in answer A_2 would thus make this answer overinformative with respect to the functional needs defined by the main question.[14] In other words, in

---

[12] By definition, the goal of a topic-constituting question is the requested final comment value to the main topic which it has introduced.

[13] The context dependence of the notion of a satisfactory (uniquely determining) answer is illustrated in Van Kuppevelt (1991) and is a central point of discussion in Van Kuppevelt (forthcoming).

[14] In this respect Kempson (1986: 96–97) provides the following account in terms of Relevance Theory (Sperber and Wilson 1986): ". . . if the utterance in question [A_2] is interpreted as involving, say 'at least two', it must be that this satisfies relevance in such a way that narrowing down the interpretation to the more precise 'no more and no less' interpretation would not increase the relevance . . . . there is no number more relevant than 'two', even if

this example the main question Q_1 gives rise to a topic weakening process implying a qualitative reduction of the original topic range associated with subquestion Q_2. Because of question Q_1, this range consisting in the set of cardinal numbers $\{1, 2, 3, \ldots\}$ is weakened to the set $\{<2, \geqslant 2\}$ (or: $\{1, 2 \vee 3 \vee \cdots\}$) representing the same set of numbers in a less specific way. From this set the satisfactory answer A_2 selects the 'at least' value $\geqslant 2$.

3. Topics and Linguistic Scales

3.1. *Introduction*

In Section 2 we gave a brief outline of the framework that accounts for discourse structure in terms of topic-forming questions. This section focuses on presenting an account of the phenomenon of scalar implicatures in terms of this theory. The presentation comprises both an account of scalar implicatures generated on the local level, i.e. those induced as the result of individual sentences for which a sentence topic is defined, and an account of scalar inferences associated with larger discourse units for which a higher-order topic or discourse topic is defined. As illustrated above, both the local and global structural levels are analyzed in a uniform way in terms of topic-forming questions.

Central to the argument in this section is the view that on each discourse level the generation of a scalar inference is determined by the explicit or implicit (sub)topic-forming question defining the particular level. First, in Subsection 3.2, we will put forward two claims in respect of these inferences, one referring to their actual generation and the other to their inferential status. Second, in Subsection 3.3, we propose a definition of linguistic scales in terms of topic-forming questions. A linguistic scale is characterized as an ordered topic range, providing scale values that share the same topic. We will illustrate how the definition provides a solution for the problems related to linguistic scales. Finally, in Subsection 3.4, we focus on the relationship between scalar inferences and discourse structure.

B [= A in our example] has a larger number of children". As is obvious from our illustration, the phenomenon of the 'at least two' interpretation being optimally relevant is explained in this context by the fact that this value is already sufficient for satisfying the higher-order question, as a consequence of which a more specific exact value would be superfluous and thereby would violate Grice's (1967/75) second submaxim of quantity which says "do not make your contribution more informative than is required (for the current purposes of the exchange)".

3.2. *Scalar Implicatures Determined by Topic-Forming Questions*

3.2.1. *The Problem of Scale Activation in Elimination Contexts*

Grice (1967/75) introduced two submaxims of quantity underlying cooperative conversation. The submaxims express the principle of being neither underinformative nor overinformative. They are formulated as follows:

The Maxim of Quantity
(i) make your contribution as informative as is required (for the current purposes of the exchange)
(ii) do not make your contribution more informative than is required

The first submaxim, the one giving rise to what is called Quantity-(i) implicatures, is central to our discussion. In this paper we focus on one particular subclass, namely that of scalar Quantity-(i) implicatures.[15]

Levinson (1983:106) explains the phenomenon of (scalar) Quantity-(i) implicatures in terms of the submaxim in question as follows:

Suppose I say:

(24) Nigel has fourteen children

I shall implicate that Nigel has only fourteen children, although it would be compatible with the truth of (24) that Nigel in fact has twenty children. I shall be taken to implicate that he has only fourteen and no more because had he twenty, by the maxim of Quantity ('say as much as is required') I should have said so. Since I haven't, I must intend to convey that Nigel only has fourteen.

In other words, in agreement with the standard interpretation of quantifying terms in first order predicate logic, the assertion that Nigel has fourteen children entails a lower bound that he has *at least* fourteen children, but pragmatically implicates an upper bound that he has *no more than* fourteen children. Semantic and pragmatic inference together result in the interpretation that Nigel has *exactly* fourteen children.[16]

Horn (1972) proposed the notion of linguistic scales underlying Quantity-(i) implicatures as conveyed in Levinson's example given above. A linguistic scale is assumed to be activated or triggered by the quantifying term *fourteen*. This scale is a (linear) ordered set of alternative cardinal

---

[15] We are conscious of the fact that another subclass of Quantity-(i) implicatures, called *clausal implicatures*, is also determined by topic-forming questions. For instance, the unsatisfactory answer *A believes p* given to the question *Who is screaming?*, whereby *p* is '(The one who is screaming is) Bill', gives rise to the clausal implicatures 'possibly *p*' and 'possibly $\neg p$'. These implicatures represent the corresponding set of possible answers $\{p, \neg p\}$ to this question.

[16] In this paper we abstract from the so-called property of the epistemic modification of pragmatic inferences (see, e.g., Levinson 1983 for a brief discussion of this subject).

numbers, e.g. the finite set ⟨16, 15, 14, 13⟩. The prediction of scalar inferences on the basis of this scale is as follows. Given a scale and the assertion of a sentence containing a value on that scale, e.g. the assertion that Nigel has fourteen children, it is implicated that sentences containing a cardinal number higher on this scale, in this case the sentences that Nigel has fifteen and sixteen children, are negated.

An essential characteristic of implicatures is that they can be denied without a logical contradiction arising. The elimination of an implicature is achieved either explicitly, by means of an additional phrase or statement, or implicitly, as the result of the linguistic or non-linguistic context. As will be discussed in the next subsection, this possibility of elimination is generally considered to be a distinguishing property which expresses the status of implicatures as a weaker type of inference than (semantic) entailment.[17]

Using the above example we can illustrate the property of elimination without logical contradiction as follows.

(6) Nigel has fourteen children. In fact he has twenty.

The standard explanation is that the first utterance gives rise to the implicature that Nigel has no more than fourteen children and that in the second utterance this implicature is explicitly cancelled.

Other types of implicature elimination given in the literature include:

(7) Nigel has fourteen children, if not more

(8) Nigel has fourteen children and maybe more.

In (7) and (8) the implicature that Nigel has no more than fourteen children is suspended both by the *if*-clause and the expression *maybe more*.

However, taking into account the topic-comment modulation of a sentence, it is at least doubtful whether an implicature is actually generated in the supposed elimination processes. Consider the following question-answer pair.

(9) Who has fourteen children?
Nigel$_{Comment}$ has fourteen children. In fact he has twenty.

Given the question, the constituent *Nigel* in the answer has comment function. According to standard implicature theory, the second utterance

[17] We assume a notion of semantic entailment as defined by Horn 1972, Gazdar 1979 and others, namely: **p** entails **q** when **q** is true under every assignment of truth values under which **p** is true.

cancels the upper bound implicature associated with the quantifying term *fourteen* in the first utterance. However, no evidence exists for the assumption that this implicature is generated at all and that it is followed by cancellation. The semantic interpretation of this sentence that Nigel has *at least* fourteen children is completely compatible with the assertion that he has twenty. It seems that in cases where the assumed implicature inducing expression is not part of the comment, implicature theory assumes, without evidence, two extra processes, namely a process of generation and one of cancellation. As will be demonstrated further in the next section, both processes are redundant given the semantic interpretation 'at least fourteen' of the lexical item of which it is assumed that it gave rise to the implicature.[18,19]

However, based on the fact that questions are induced as the result of indeterminacies (Section 2.2), it is not only doubtful but even highly unlikely that a scalar implicature is generated at all in these cases. If in (9) a scalar implicature would have been induced as the result of the quantifying term *fourteen*, this would transform the semantically provided 'at least fourteen' interpretation of this term into 'exactly fourteen' implying that this term is no longer an indeterminacy and that, as a consequence, question induction is blocked. Example (9)' illustrates that this prediction is wrong.

(9)' Who has fourteen children?
Nigel$_{Comment}$ has fourteen children.
⟨How many children does he have?⟩
He has twenty$_{Comment}$.

The answer that Nigel has fourteen children gives rise to the (implicit) question asking for the exact number. By definition, this question is only induced if the quantifying term *fourteen* is an indeterminacy the interpretation of which is 'at least fourteen' rather than 'exactly fourteen'.

We will argue that a different implicature is induced in (9): not the quantifying term *fourteen* but the term *Nigel* gives rise to an implicature

---

[18] A quantifying term like *fourteen* thus receives a monotone increasing interpretation if it has no comment status: adding more referents than fourteen does not change the truth value of the sentence containing this expression. In this paper we abstract from cases in which such a quantifying term represents an old comment value, implying that it got an 'exactly *n*' interpretation in the preceding context.

[19] In this respect Kroch (1972) already noted that the notion of conversational implicature gives rise to a problem of overgeneration due to vague concepts which are difficult to test.

because it has comment status.[20] It denies all other candidates in the given context as the persons who have (at least) fourteen children. In other words, one of our criteria for implicature generation, and thus for scale activation (see also Section 3.4.1), is that the inducing context must have comment function.[21]

The importance of comment function for the generation of scalar implicatures, especially those induced as the result of cardinal numbers, has also been observed in Campbell (1981) and, extensively, in Fretheim (1992). According to Campbell, only when cardinals are in comment position do they always give rise to an 'exactly *n*' interpretation. However, in contrast to what is argued in the next section, it is assumed that this interpretation involves a pragmatic rather than a semantic inference of the type 'no more than *n*'. According to Fretheim, on the other hand, cardinals within the 'focus domain' (i.e. those that are part of the comment) get an 'exactly *n*' interpretation of which the upper bound is given not by conversational implicature but is part of the linguistic meaning of the utterance itself. Only if a cardinal belongs to the background (topic part), is its upper bound pragmatically provided by conversational implicature. Apart from the fact that no evidence exists for the assumption that in the latter case an inference is actually generated, the preceding example (9)′ illustrates that this possibility is ruled out by the simple fact that in such a case question induction is still an option. Furthermore, Fretheim leaves open the question as to whether there is also a linguistic scale activated in cases in which the cardinal is part of the comment. If so, it is not at all clear how such a scale differs from one activated by cardinals that are not in comment position. In the former case, as will be argued

---

[20] As may become clear from our analysis of linguistic scales (Section 3.3.2) we do not adopt the position in respect of quantifying terms that they give rise to scalar inferences only if they do *not* belong to the comment part of the sentence (see, e.g., Seuren 1993).

[21] Another strong argument supporting the view that the inducing context of inference must have comment function can be derived from the property of *reinforceability* (Sadock 1978, Levinson 1983), which says that conversational implicatures can be explicitly added to the inducing context without causing an unacceptable redundancy. Consider the following two question-answer pairs.

\# How many cookies did Billy eat?
Billy ate three cookies but not all.

Who ate three cookies?
Billy ate three cookies, but not all.

An (unacceptable) redundancy seems to occur only where an upper bound inference is actually generated, as is the case in the first pair. No such inference is generated in the second pair. This implies that the assertion that Billy did not eat all cookies does not have a redundant status.

for in the next section, the scalar inference must be an entailment in agreement with the 'exactly *n*' meaning of the cardinal, while in the latter case this inference is a weaker pragmatic one that can be cancelled. Furthermore, assuming that a linguistic scale is also activated in the case in which the cardinal belongs to the comment, it is unclear how such a scale has to be defined, when we take into account that its elements constitute exact, mutually exclusive values that themselves exclude the one-sided ordering relation characteristic for linguistic scales.

The criterion for the generation of an upper bound scalar inference, namely that the inducing context must have comment function, obviously presupposes that the provided comment value is one that does not comprise the highest value on the associated linguistic scale. Satisfactory answers which provide an 'exact' value therefore usually give rise to an upper bound scalar inference. This is certainly the case when the 'exact' value does not constitute the end value on the associated linguistic scale. On the other hand, unsatisfactory answers that merely provide a less specific 'at least' value do not give rise to such an inference, since the higher end value on the linguistic scale is included in the 'at least' value. However, as we will illustrate, in one specific case an upper bound scalar inference may also be induced as the result of an unsatisfactory answer, namely if such an answer merely selects a subrange of scale values excluding higher scale values.

The criterion, therefore, also accounts for the phenomenon of topic weakening by higher-order questions, as was illustrated above by Kempson's (1986) example which is repeated here as example (10).

(10)
F_1	A:	I'm a mother.
Q_1		Do I get a fixed amount of state benefit?
A_1	B:	If you have at least two children, you get a fixed amount of state benefit.
$\langle Q_2 \rangle$		⟨How many children do you have?⟩
A_2	A:	I have two children. (In fact I have four.)

We argued that a topic weakening process is involved in this example, implying that the original topic range of subquestion Q_2, i.e. $\rho(T_2) = \{1, 2, 3, \ldots\}$, is reduced to $\rho'(T_2) = \{<2, \geqslant 2\}$ as the result of the main topic-constituting question Q_1 to which subquestion Q_2 is subservient. Because of the relationship between topic ranges and linguistic scales, a relationship that will be accounted for in detail in Section 3.3, the linguistic scale associated with the comment value in answer A_2 is not $\langle \ldots, 3, 2, 1 \rangle$ but the reduced linguistic scale $\langle \geqslant 2, <2 \rangle$. However, in this case no upper bound scalar implicature is induced, because the comment value provided

by the satisfactory answer A_2, i.e. the value 'at least two' ('≥2'), constitutes the highest value on this scale. Although no upper bound inference is generated, answer A_2 entails the corresponding sentence containing the lower value on this scale, namely if speaker A has at least two children, i.e. two, three or more, A also has less than two children, namely one.

We already said that the criterion implies that no upper bound scalar implicature is generated as the result of unsatisfactory answers providing an 'at least' value. In this respect we distinguish two different situations of an unsatisfactory answer, namely one in which the unsatisfactory answer is completed by means of a process of subquestioning and one in which a satisfactory answer cannot be given because of an epistimic limitation.[22] Consider first the following analysis of an example presented by Horn (1992: 175), which illustrates the former situation.

(11) Q_1 A: ⟨How many of your friends are linguists?⟩
Are many$_{Comment}$ of your friends linguists?
A_1 B: Yes,
⟨Q_2⟩ ⟨How many?⟩
A_2 (In fact) all$_{Comment}$ of them.

In (11) answer A_1 is unsatisfactory, given the succeeding extension A_2 which is considered to be realized by means of the implicit subquestion Q_2. The comment value *many* which is confirmed by this answer must be interpreted as 'at least many', thereby comprising a set of higher 'exact' values on the corresponding linguistic scale. Given that the linguistic scale ⟨all, most, many, some, few⟩ is activated, the unsatisfactory comment value implies that a satisfactory answer to question Q_1 is either '(exactly) many', '(exactly) most' or '(exactly) all', implying that only the lower scale values *some* and *few* are excluded as possible ('exact') answers to this question. Because the unsatisfactory answer provides an 'at least many' comment value, no higher value is left as possible satisfactory answer to the question and, as a consequence, no upper bound scalar inference is generated. In Section 3.2.3 we discuss in greater detail the 'at least many'–'exactly many' distinction and comparable distinctions. In addition, we provide a criterion for their identification.

As stated in Section 2.2, an answer may be unsatisfactory because of an epistemic limitation, namely if the knowledge of the addressee fails to provide a satisfactory answer to the question.

---

[22] We will not discuss the situation in which the question process is broken off or disturbed in some other way.

(12) Q_1 A: Which of the boys John, Andy and Martin went to the match?
A_1 B: All I know is that John and Andy went.

In (12) the provided comment value is 'at least John and Andy', which does not exclude the implied, stronger value 'John, Andy and Martin' as a possible (exact) answer to Q_1. If answer A_1 had been an exact answer, the latter value would have been excluded by means of an upper bound inference. An account of examples like these in terms of linguistic scales is presented in Subsection 3.3.3.

In very specific cases, answers that are unsatisfactory because of an epistemic limitation give rise to upper bound scalar inferences. But this happens only if the unsatisfactory comment value is not an 'at least' value. Consider the following two related examples.

(13) How many children does Nigel have?
Nigel has fourteen$_{Comment}$ children, if not fifteen$_{Comment}$.

(14) How many children does Nigel have?
Nigel has fourteen$_{Comment}$ children and maybe fifteen$_{Comment}$.

Two points have to be mentioned in this respect. First, both answers are unsatisfactory because of an epistemic limitation, but they do not provide an 'at least' comment value. Therefore, they give rise to an upper bound scalar inference. Second, only the whole answers *fourteen [. . .], if not fifteen* and *fourteen and maybe fifteen* give rise to an upper bound scalar implicature, namely one excluding sentences containing scale values higher than fifteen. This is contrary to implicature theory which assumes an elimination process consisting of the suspension of the supposed implicature triggered by *fourteen*. However, this value is only *part* of the comment value provided by the sentence. As in the case of (9) there is no evidence for the assumption of two extra processes of generation and elimination of an implicature.

We can conclude here that upper bound scalar inferences while they are generated as the result of comment values, this is only if these values do not comprise the highest value on the associated linguistic scale. Usually this is the case with satisfactory answers, because they provide an 'exact' comment value. Unsatisfactory answers mostly contain an 'at least' value. In such cases the generation of an upper bound inference is impossible.

3.2.2. *Entailments Instead of Weaker Pragmatic Inferences*

3.2.2.1. *The 'Elimination by Contradiction' Cases.* Horn (1972) distinguishes two ways of eliminating a scalar Quantity-(i) implicature: 'elimination by contradiction' and 'elimination by suspension'. In the former case the implicature is (explicitly) cancelled, implying that its truth value is denied. An illustration of this phenomenon is example (6) given above. In the latter case the implicature is eliminated on grounds other than its truth value, namely by 'explicitly leaving the possibility open that a higher value on the relevant scale obtains' (Horn 1989: 235). Examples of the latter type of elimination are (7) and (8). In this and the next subsection we will argue that what is considered to be an essential characteristic of scalar Quantity-(i) implicatures, namely that they can be eliminated without giving rise to a contradiction, does not hold in every situation. This supports the view that inferences of this type are in fact (semantic) entailments generated in specific contexts rather than weaker pragmatic inferences. The latter can, in principle, be induced in all contexts, though they may be the subject of explicit or implicit elimination.

It generally holds that if no correction is involved a semantic entailment cannot be cancelled without a contradiction arising.

(15) #Harry bought four books. In fact he bought three.

The second utterance in (15) contradicts the first one, both in its 'exactly four' interpretation which it has if the quantifying term *four* has comment status and in its 'at least four' interpretation if this term belongs to the topic part of the utterance.

In (16), on the other hand, no contradiction is involved in the cancellation process. In this case the cancellation is a correction, implying the replacement of one value by another.

(16) A: Harry bought four books.
B: No, he bought seven.

But now consider the examples (17) and (18).

(17) Who bought four books?
Harry$_{Comment}$ bought four books. In fact he bought seven.

(18) [Harry did a lot of shopping this afternoon.]
How many books did he buy?
#He bought four$_{Comment}$ books. In fact he bought seven.

In contrast to example (17), the quantifying term *four* in (18) has comment status. As was argued for in the preceding subsection, it must give rise to

an upper bound scalar inference, namely that Harry bought *no more than* four books. Evidence for the fact that this inference is actually generated is that if we assume the opposite the answer could not have been unacceptable. This is the case if a topic weakening process would have been involved, like in the following variant of (18).

(18)′ [Did Harry get a free book in this shop?
If he bought four books, he got one.]
How many books did he buy?
√He bought four$_{Comment}$ books. In fact he bought seven.

As in (17), the assertion that Harry in fact bought seven books is compatible with the lower bound interpretation of *four* in the first part of the answer.

The unacceptability of the answer in (18) demonstrates two points. First, it demonstrates that the cancellation of an upper bound inference cannot, as is generally assumed, be an essential property of scalar Quantity-(i) implicatures. Example (18) shows that the cancellation of such an inference may be blocked. Second, it demonstrates that if cancellation is not possible it is blocked by a contradiction which implies that the upper bound inference that Harry bought no more than four books is *entailed* by the first part of the answer.

In (18) two types of entailments are associated with the quantifying term *four*, the meaning of which is *exactly four*. The answer in which this term occurs entails sentences containing a lower value on the corresponding linguistic scale of cardinal numbers. However, because of the scale ordering, these values must be interpreted as 'at least' values instead of 'exact', mutually exclusive ones. If the number of books Harry bought is (exactly) four, it follows that he bought (at least) three or less books. Besides these entailments, the answer also entails the negation of sentences containing a value higher on the scale, e.g. the negation of the sentence that Harry bought (at least/exactly) five books. Inferences of the latter type are upper bound scalar inferences which, in our opinion, are incorrectly characterized in pragmatic theory as a weaker type of inference. They differ from lower value inferences in the fact that they cannot can be inferred in all circumstances independent of the topic-comment modulation of the sentence. Both *Harry bought four*$_{Comment}$ *books* and *Harry*$_{Comment}$ *bought four books* entail that he bought (at least) three, two, etc. books, but only the former statement entails that he did not buy more than four books.

The phenomenon described here can also be demonstrated by yes/no-questions with an underlying WH-question.[23,24]

(19) [Who bought four books?]
Did Harry$_{Comment}$ buy four books?
Yes, in fact he bought five.

(20) [Harry bought a lot of books.
How many books did he buy?]
Did he buy four$_{Comment}$ books?
#Yes, (in fact) he bought five.

The answer in (20) is unacceptable because, in contrast to the answer in (19), it involves a contradiction due to the fact that, first, an upper bound inference is induced as the result of the confirmation of the comment value *four* in the question and, second, this inference is an entailment.[25]

Summarizing so far we can say that an upper bound scalar inference in fact is a context dependent inference that is not induced under all topic-comment modulations of the inference inducing sentence. This inference is an entailment and not a weaker pragmatic inference which can be cancelled without causing a contradiction.

---

[23] We do not adopt Campbell's (1981) distinction between phenic (effortful) and cryptic (automatic, effortless) processes in identifying the generation of a scalar inference. Whether an upper bound scalar inference is generated at all in the case of, e.g., the yes/no-question *Did Harry buy four books?* is considered to be a phenic, effortful process, while whether one is generated in case of the answer *Four* given to the corresponding WH-question *How many books did Harry buy?* is considered to be a cryptic process involving no real effort on the part of the questioner. According to our criterion a scalar inference is generated if the inference inducing element has comment function. However, in discourse this function is intonationally and/or contextually marked, implying that in principle no extra effort is needed to determine the existence of a scalar inference.

[24] This type of yes/no-question is intonationally and syntactically accounted for elsewhere (Van Kuppevelt 1991).

[25] The observation that context plays a crucial role in the actual generation of inferences of this type is anticipated in Horn (1972) and also observed in Scharten (forthcoming). Horn (1972: 33):

(1.63)a. Does John have three children?
b. Yes, (in fact) he has four.
c. No, he has four.

As in the case of the examples (17), (18) and (18)' the yes/no-question is, in our view, ambiguous between one that underlies the question *Who has three children?* and one that underlies the question *How many children does John have?* This ambiguity is intonationally marked by the main accent on different constituents in the question, namely on *John* and *three* respectively.

3.2.2.2. *The 'Elimination by Suspension' Cases.* A special category of constructions causing the elimination of a scalar Quantity-(i) implicature are the suspension cases. The examples given include those of the following syntactic form (see Horn 1972/89).

(21) (at least) $\mathbf{P_i}$, if not (downright) $\mathbf{P_j}$
$\mathbf{P_i}$, {or, and possibly} even $\mathbf{P_j}$
not even $\mathbf{P_i}$, {let alone/much less} $\mathbf{P_j}$
($\mathbf{P_j}$ and $\mathbf{P_i}$ are elements of the same linguistic scale such that $\mathbf{P_j} > \mathbf{P_i}$)

According to the standard explanation the suspending clauses containing the scale values $\mathbf{P_j}$ eliminate the implicatures induced as the result of the lower scale values $\mathbf{P_i}$. Though the final inferential outcome is that in fact no inference is induced as the result of $\mathbf{P_i}$, there is no evidence for the assumed generation and cancellation processes as we have said above (Section 3.2.1). However, as already indicated by Horn (1989), there is another reason that supports the view that such an assumption is redundant. The value $\mathbf{P_i}$ which it is assumed gives rise to an implicature is not a stable one as compared to other values which certainly give rise to an implicature. This value is merely a part of the underdetermined, non-unique value expressed by the construction as a whole. In the context of the suspension construction *P_i, if not P_j*, this means that either $\mathbf{P_i}$ or $\mathbf{P_j}$ is the case. In other words, as long as it is not known whether $\mathbf{P_i}$ is the case, it is unlikely that an inference would be induced as the result of it.

Characteristic for the elimination by suspension cases is that the whole and not just a part of the construction constitutes an answer to a topic-forming explicit or implicit question. Consider the following two question-answer pairs.

(22) How many books did Harry buy?
Harry bought four$_{Comment}$ books, if not five$_{Comment}$.

(23) [Someone of my group bought no less than four books.]
Who bought four books?
Harry$_{Comment}$ bought four books, if not five.

Compared to (23) the acceptability of the answer in (22) demonstrates that the suspending clause *if not five* is part of the underdetermined, non-unique comment value provided by the whole answer.[26]

The view that the whole suspension construction and not just a part of

[26] We abstract here from so-called echo questions like *Who bought four if not five book?*

it constitutes the comment is further supported by the phenomenon of multi-comment sentences. Consider in this respect example (23)′ which, in contrast to (23), is an acceptable (part of) discourse.

(23)′ [I would like to know who bought how many books.]
Who bought $\underline{\text{four}}_{Comment}$ books?
$\surd\underline{\text{Harry}}_{Comment}$ bought $\underline{\text{four}}_{Comment}$ books, $\underline{\text{if not five}}_{Comment}$.

Suspension constructions must, therefore, be distinguished from constructions such as *many . . . in fact all* which cannot be interpreted as constituting one comment value. As shown by our preceding example (11), these constructions comprise an intervening explicit or implicit subquestion, implying that two – completing – comment values are involved.

If a suspension construction gives rise to a scalar inference at all, it is generated as the result of the whole construction ('$\mathbf{P_i}$, if not $\mathbf{P_j}$'), and not just the non-suspending part of it ('$\mathbf{P_i}$'). The status of the latter is not a definite answer providing a unique comment value, but is merely an answer providing a possible comment value out of the range of comment values defined by the whole answer given to the question, e.g. the range of possible comment values {4, 5} in the case of example (22).

Suspension constructions as a whole give rise to an implicature, only if the upper bound of the selected range of possible comment values lies before the upper bound of the whole range associated with the question. For instance, if in (22) the question were to define the finite (topic) range {4, 5, 6}, the answer which selects the subrange {4, 5} would give rise to the implicature that Harry did not buy six books.

3.2.2.3. *Gazdar's (1979) Arguments.* Gazdar (1979) offers two arguments to support the view that scalar Quantity-(i) implicatures cannot be entailments. We will demonstrate that neither argument can be upheld when viewed contextually, particularly not if we take into consideration the topic-forming explicit or implicit question answered by the inference inducing sentence.

The first argument, which we call the *cancellation argument*, has in fact been discussed earlier in another way. Consider the following set of sentences.

(24) a. Some of the boys were at the party.
b. Not all of the boys were at the party.
c. Some, in fact all, of the boys were at the party.

Gazdar's cancellation argument goes as follows: the relation between the a-sentence and the b-sentence cannot be an entailment, for entailments cannot be cancelled (without contradiction) and the c-sentence shows that cancellation of the inference is quite possible. However, the argument cannot be sustained if the quantifying terms in the a-, b- and c-sentence occur in comment position. In such a case the quantifying term *some* in the c-sentence does not give rise to a scalar inference as in the a-sentence and, consequently, no cancellation process is involved. In the c-sentence this term does not constitute an 'exact' value but an unsatisfactory 'at least' value which is completed by the *in fact all* phrase. An analysis of this situation is given in (24)'.

(24)'c. How many of the boys were at the party?
Some, ⟨how many?⟩ in fact all, of the boys were at the party.

But now consider the following set of sentences relevant to Gazdar's second argument. We will call this argument the *inconsistency argument*.

(25)a. Some of the boys were at the party.
b. Not all of the boys were at the party.
c. All of the boys were at the party.

In this case the argument is one of reduction ad absurdum. The c-sentence entails the a-sentence. If the a-sentence would entail the b-sentence, then (by transitivity) the c-sentence would entail the b-sentence which, of course, cannot be the case since the one sentence contradicts the other. Therefore, the a-sentence cannot entail the b-sentence.

The argument becomes invalid if we take into account the topic-comment modulation of the sentences. Let us first suppose that in each sentence the quantifying value in sentence initial position has comment function, e.g. by considering it to be an answer to the topic-forming question *How many of the boys were at the party?* There are (at least) two situations in which this argument becomes invalid. First, the entailment relation between the c-sentence and the a-sentence does not hold if all values are interpreted as 'exact' answer values: if the number of boys at the party were *all* of them, then this number is not *some* of them. Second, as is generally assumed, the c-sentence entails the a-sentence, implying the following: if the number of boys that were at the party were *all* of them, then this number is *at least some* of them. However, this non-exact value in the a-sentence implies that this sentence does not entail the b-sentence.

If, on the other hand, in each sentence the constituent *at the party* has comment function, no evidence is available to support an inference

relation between the a-sentence and the b-sentence. We have already seen that no cancellation process involving such an inference takes place in the following cases.

(26)a. Where were some of the boys?
b. Some of the boys were at the party, in fact all.

The 'at least' interpretation of the quantifying term *some* in the answer comprises the possible exact answer value *all*, as a consequence of which the inference between the a-sentence and the b-sentence is blocked.

3.2.3. *Extending the Analysis to Non-Cardinals*

Central to the preceding subsections was the hypothesis relating to the topic dependency of the generation of upper bound scalar inferences. We saw that these are induced as the result of sentence parts representing comment values, but only of those parts which do not comprise the highest value on the associated linguistic scale. Therefore, comment values which represent an unsatisfactory 'at least' value, i.e. those which are completed as the result of subquestions as well as those which involve an epistemic limitation, do not give rise to such an inference. Furthermore, comment values constituting a satisfactory answer that comprises the highest scale value are also excluded from generating such an inference. However, in the preceding subsections we have concentrated mainly on cardinals. In this subsection we show that the hypothesis also relates to upper bound scalar inferences induced as the result of non-cardinals, whereby we especially pay attention to the problems observed in the literature. Consider the following examples.

(27) ⟨*all, most, many/much, some*⟩
a. Q_1 How much of the profit does John get?
A_1 Much$_{Comment}$ of it. ⟨How much?⟩ In fact most$_{Comment}$.
b. Q_1 Who got much of the profit?
A_1 John$_{Comment}$. ⟨How much?⟩ In fact he got most$_{Comment}$.

(28) ⟨*and, or*⟩
a. Q_1 What would you like to drink? Would you like tea or coffee$_{Comment}$?
A_1 Yes, ⟨Which of the two?⟩ both$_{Comment}$ please.
b. Q_1 Who would like tea or coffee?
A_1 John$_{Comment}$. ⟨Which of the two?⟩ In fact he would like both$_{Comment}$.

(29) ⟨*certain, almost certain, pretty likely, likely, possible*⟩

a. Q_1 How likely is it that John will decide to enrol for this major program?
   A_1 Pretty likely$_{Comment}$. ⟨How likely?⟩ In fact it is almost certain$_{Comment}$.

b. Q_1 Of which student is it pretty likely that he will decide to enrol for this major program?
   A_1 It is pretty likely that John$_{Comment}$ will decide to enrol for this major program. ⟨How likely?⟩ In fact it is almost certain$_{Comment}$.

In the answers of all the a-parts an unsatisfactory comment value is completed to a satisfactory one by means of an (implicit) subquestion. The hypothesis predicts that no scalar inference is induced as the result of the unsatisfactory answers, but solely as the result of the satisfactory ones. However, as is also predicted by the hypothesis, the satisfactory answer 'both' in (28a) forms an exception, because it represents the highest value on the associated linguistic scale.

As is the case with the unsatisfactory answers in the a-parts, the values *much*, *tea or coffee* and *pretty likely* in the b-parts also represent an 'at least' value and, as predicted, do not give rise to an upper bound scalar inference. However, unlike the unsatisfactory answers in the a-parts, they represent an 'at least' value because of the fact that they do not have comment status.

In contrast to our position, Carston (1985/88) provides an analysis of cardinals which implies that cardinals must be distinguished from non-cardinals. The author argues against the standard 'at least' semantics of cardinals in favor of a so-called neutral semantics. According to Carston, the linguistic meaning of cardinal predicates is not 'at least *n*'. Depending on the context, i.e. both the linguistic context and the non-linguistic context containing background knowledge, these predicates get an 'at least *n*', 'at most *n*' or an 'exactly *n*' interpretation. Carston's view is supported by, among others, the following examples which imply, respectively, a standard, non-linguistically determined 'at least' and 'at most' interpretation.

(30) You don't have to be (at least) sixteen to drive a car; you have to be (at least) eighteen.

(31) She can have (at most) 2000 calories a day without putting on weight.

However, as demonstrated by the following variants analyzed in terms of

question–answer structure, an upper bound scalar inference is also generated in these cases, depending on the question answered.

(30)′a. Q_1 How old do you have to be to drive a car in Holland?
A_1 In Holland you have to be eighteen to drive a car.
b. Q_1 In which country do you have to be sixteen to drive a car?
A_1 In Holland. ⟨How old then?⟩ You have to be eighteen.

(31)′a. Q_1 How many calories a day can Jane have without putting on weight?
A_1 Jane can have 2000 calories a day without putting on weight.
b. Q_1 Who can have 2000 calories a day without putting on weight?
A_1 Jane. ⟨How many can she have?⟩ In fact she can have 2800 calories a day without putting on weight.

Only the answers in the a-parts give rise to an upper bound scalar inference. In contrast to those in the b-parts they contain a cardinal in comment position. The scalar inference induced as the result of the answer in (30a)′ is that in Holland you don't have to be (at least) nineteen or older to drive a car. The one generated as the result of the answer in (31a)′ is that Jane cannot have, e.g., (at most) 2500 calories a day without putting on weight.

Apart from these examples, Carston further underpins her claim with examples in which the 'at least'/'at most' interpretation is not standardly given by background knowledge, but by contextual factors. Consider the following example presented by Carston.

(32) If Mrs. Smith has no more than three children, we'll all fit into the car. She does have (at most) three children.

Obviously, as in our preceding example (5), a topic weakening process is involved in this example. The linguistic scale associated with this example is ⟨>3, ⩽3⟩. Apart from the fact that the cardinal in the second sentence receives an 'at most' interpretation, it also gives rise to the generation of an upper bound scalar inference, namely that Mrs. Smith does not have more than three children.

We can conclude that Carston's observations do not form an exception to our central hypothesis relating to the generation of upper bound scalar inferences. In this respect the 'at least *n*'–'exactly *n*' distinction can, in principle, be preserved. As demonstrated by the examples (27), (28) and (29), the 'at least *n*' interpretation is obtained if the cardinal is not in

comment position, while the 'exactly *n*' interpretation holds if it has comment status. However, as may be clear from these and previous examples, this distinction has to be interpreted in terms of un(der)determinedness and determinedness respectively.

Apart from Carston's analysis, other arguments are given in the literature to account for the special status of cardinals. Most of these arguments are discussed in Horn (1992) who provides, among other things, a defense of the standard theory of scalar implicatures. We will discuss briefly some of these arguments and demonstrate that the behavior of cardinals and non-cardinals is predicted in a uniform way by our hypothesis.

As Sadock (1984) has argued the truth conditions of mathematical statements require that the cardinals involved have 'exact' meanings rather than 'at least' meanings as is implied by standard implicature theory. The problem is that a mathematical statement such as, '2 + 2 = 3' would be true if the cardinal '3' would mean 'at least 3'. However, the behavior of the arguments of two-place predicates like '+' and '=' involved in mathematical statements is in agreement with cardinals, in ordinary statements when they are in comment position. These arguments typically express a determined, specified quantity like cardinal numbers in comment position, e.g. the cardinal number *three* in *Edgar has three cars* as an answer to the question *How many cars does Edgar have?* For that reason, a mathematical statement such as 2 + 2 = 4 must be analyzed as an answer to the three-fold question *How much plus how much equals how much?* or, depending on the context, as an answer to, e.g., the reduced one-fold question *How much is* $\underline{2}_{Comment}$ *plus* $\underline{2}_{Comment}$*?*

Similar to cardinals in mathematical statements, those that are lexically incorporated, such as the cardinals in the words 'triangle' and 'quadrangle', always have an 'exactly *n*' meaning rather than an 'at least *n*' meaning (e.g., Horn 1972 and Hirschberg 1985). For example, the cardinal *three* incorporated in 'triangle' specifies the number of angles of the denoted object, thereby defining uniquely this geometric figure. For this reason, this cardinal must be interpreted as answering an incorporated 'how many'-question and, accordingly, 'triangle' must be interpreted as 'how many'-angled figure'.

As Horn (1972) and others have argued, cardinals constituting an approximative value are more likely to be interpreted as non-inference inducing 'at least' values, e.g. as is the case in 'John has $200' as compared to 'John has $201.37'. It is said that this phenomenon does not occur with non-cardinals. However, our hypothesis implies that approximations do not have a special status and that they answer the general rule of implicat-

ure generation. If in comment position, an approximative value is either an unsatisfactory answer completed by a process of subquestioning to a satisfactory, more precise, value or a satisfactory answer resulting from a process of topic weakening.

A remarkable observation concerns the relationship between collective and distributive readings on the one hand, and 'exactly *n*' and 'at least *n*' meanings on the other hand (e.g., Atlas 1990 and Horn 1992). Compare the following two sentences presented in Horn (1992).

(33)a. If there are three books by Chomsky (in the shop), I'll buy them all.
b. If there are three books by Chomsky (in the shop), I'll buy each of them.

In the a-sentence, which involves a collective reading, the cardinal *three* is interpreted as 'exactly three', while in the b-sentence, which represents a distributive reading, this cardinal has an 'at least' meaning. As shown in (33)', this observation is fully explained by our hypothesis which implies that cardinals are not distinct from non-cardinals in this respect.

(33)'a. Q_1 How many books by Chomsky will you buy'?
A_1 If there are three books by Chomsky (in the shop), I'll buy them all.
b. Q_1 Which books will you buy?
A_1 If there are three books by Chomsky (in the shop), I'll buy each of them.

In contrast to answer A_1 in the a-part, A_1 in the b-part cannot be interpreted as answering a 'how many'-question. So, only in the former case the cardinal represents a comment value, and for that reason is interpreted as 'exactly three'.

The last argument to be mentioned here – that cardinals in comment ('focus') position are given a purely semantically provided 'exactly *n*' interpretation – has been extensively discussed in previous sections, namely as has been defended, in particular, by Fretheim (1992). We have already explained how this analysis differs from our own analysis. However, as Horn (1992) has argued, Fretheim's position would imply that cardinals should be distinguished from non-cardinals which, in comment position, may also receive an 'at least' interpretation. It follows from our hypothesis that, in this respect as well, cardinals do not have a special status. Like non-cardinals, cardinals in comment position may be given an 'at least' interpretation. This was demonstrated by Kempson's example (see example (5)) which we analyzed as a topic weakening process. On

the other hand, as has also been shown, the 'at least' interpretation of non-cardinals that have comment status is the result of unsatisfactory answers. This implies that non-cardinals, e.g. the quantifying terms *some*, *many* and *most*, may receive two interpretations in discourse, namely an 'at least' interpretation and 'exact' interpretation. An illustration of this phenomenon is the preceding example (27) in which the answer containing the quantifying term '(at least) much' is completed to one containing '(exactly) most'. Only the latter gives rise to an upper bound scalar inference which is expressed by a corresponding sentence containing the value 'not all'. If no epistemic limitation is involved, quantifying terms such as 'many' must be interpreted as 'exactly many', unless the less specific interpretation 'at least many' does not imply underinformativeness, as is the case if an answer is completed by means of subquestioning or its topic reduced by a process of topic weakening.

Once again we conclude that there is no principle difference between the behavior of cardinals and non-cardinals with respect to the central hypothesis concerning the generation of scalar inferences. If no weakening process is involved, cardinals differ from non-cardinals only in the fact that, without explicit mentioning, they always represent an 'exact' value when functioning as a comment.

3.3. *Linguistic Scales as Ordered Topic Ranges*

In Section 3.2 we discussed two characteristic properties of scalar Quantity-(i) implicatures and we argued for topical restrictions on their generation as well as for an essential change of their inferential status. It was shown first that the actual generation of an implicature of this type depends on the topic-forming explicit or implicit question that is answered in the inducing context and, second, that such a topic-based inference is an entailment and not, as is generally assumed, a weaker pragmatic inference. Obviously, both the topical restrictions and the assignment of a different inferential status affect the nature of this type of pragmatic inference.

In this section attention will be directed to the topical basis of linguistic scales underlying inferences of this type. First, we will discuss three problems that have blocked the formulation of an adequate definition of linguistic scales, namely that of *scale ordering*, *scale coherence* and *scale reduction*. After that, we will propose a definition of linguistic scales in terms of the question-based topic notion given above. This definition supplies an integrated solution to these issues, including a solution to the problem of scale activation that underlies the determination of topical restrictions on scalar inferences.

3.3.1. *The Problem of Scale Definition*

3.3.1.1. *Scale Ordering and Scale Coherence.* Central to the discussion of a definition of linguistic scales are the generally acknowledged problems of *scale ordering* and *scale coherence*. The absence of an adequate solution to these problems has led many authors (e.g., Horn 1972, Gazdar 1979, Levinson 1983) to adapt a position of taking linguistic scales as *given*, coherent sets of values ordered by semantic entailment.

Obviously, all the authors consider some ordering to be a necessary condition for linguistic scales. However, as is pointed out by Fauconnier (1975a), Hirschberg (1985) and Horn (1989), this ordering does not necessarily have to be an entailment relation. Consider in this respect Fauconnier's *pragmatic scales*. As in cases in which a scale ordering is defined by entailment, the upper bound inferences based on pragmatic scales are also entailments.

(34) A: What is the heaviest weight Alexei can lift?
B: Alexei can lift a weight of 80 pounds.

If the heaviest weight Alexei can lift is 80 pounds, it follows by entailment that he cannot lift a weight higher on the associated scale. However, no entailment relation exists between sentences containing a value on this scale. The sentence that Alexei can lift a weight of 80 pounds does not entail that he can lift a lighter one, e.g. because of its shape, form, etc.

Or, consider Hirschberg's (1985) example.

(35) A: Did you get Paul Newman's autograph?
B: I got Joanne Woodward's.

In the appropriate context B's answer entails that he did not get the autograph of a famous person higher on the scale than Joanne Woodward, thus excluding the given alternative Paul Newman. But the sentence that B got Joanne Woodward's autograph does, obviously, not entail that he actually also got one of a less famous person.

Other illustrations supporting the view that entailment cannot be the only ordering principle for linguistic scales are cited in Horn (1972/89). Among others, he gives the following illustration containing the suspension construction P_i, *if not* P_j.

(36) In the Netherlands the crowds [for the Pope] were small, the welcome lukewarm if not cold. (*New York Times*, 19 May 1985)

Apart from the fact that more ordering principles than entailment must

be assumed for linguistic scales, a general ordering criterion underlying all of them is still missing. In addition, ordering as a necessary condition for linguistic scales is clearly not a sufficient condition. Among the problems discussed in the literature are those which may be called the *hierachy problem* of scales, the problem of *scale overlap*, *scale direction* and *scale partitioning*. All the problems make it clear that an additional criterion for linguistic scales defining the set of scale values as coherent is needed.

The hierarchy problem deals with the part-whole relationship between elements on a linguistic scale. According to Hirschberg (1985) this relationship must be salient in discourse in order to give rise to scalar implicatures. Consider in this respect the following example.

(37) Q_1 A: Did Bill eat all the cake?
A_1 B: (No,) he ate some of it.

Given that the associated linguistic scale is ⟨all, . . . , some, . . .⟩, answer A_1 gives rise to the upper bound inference that Bill did not eat all the cake.

However, while salience is clearly a necessary condition it is not a sufficient one. Though ⟨all, . . . , some, . . .⟩ is an acceptable linguistic scale, ⟨. . . , Amsterdam, Holland, . . .⟩ or ⟨. . . ,tulips, flowers, . . .⟩ may not be considered a linguistic scale, despite the existence of one-sided entailment relations between the elements involved. Consider example (38).

(38) Q_1 A: Does Ed live in Amsterdam?
A_1 B: (I know) he lives in Holland.

Answer A_1 is unsatisfactory, defining a range of possible answers of which Amsterdam is one, as the result of which no scalar inference is induced. Obviously, an epistemic limitation is involved in this example.[27] But consider (38)' in which this possibility is excluded.

(38)′ Q_1 A: Do you live in Amsterdam?
A_1 B: I live in Holland.

Also in this case no scalar inference is generated, as appears from the fact

---

[27] As will become clear later, our prediction with regard to (38) is in agreement with Grice (1975: 51–52) who discusses a comparable example. Grice argues that in the case of an epistemic limitation the first submaxim of Quantity, implying not to be underinformative, is violated in favor of the maxim of Quality saying 'Don't say what you lack adequate evidence for'. In other words, the prediction with respect to the answer in (38) is that it implicates that B does not know in which town Ed is living.

that a continuation of the discourse such as *Do you live in Amsterdam, or elsewhere?* is quite acceptable.

The linguistic scales ⟨warm, lukewarm⟩ and ⟨cold, lukewarm⟩ partially overlap, leaving open the criterion of what exactly combines with the value *lukewarm* in a given context. The point is illustrated in Horn (1989: 547, n.25).

(39) My beer is lukewarm, if not downright {warm/#cool}
My coffee is lukewarm, if not downright {cold/#hot}

The problem of scale direction applies to linguistic scales which do not differ in the set of values they provide but in the *direction* of the one-sided ordering relation defined on this set. Consider example (40).

(40) a. Pete drinks five whiskies a day.
b. Pete cut down his drinking to five whiskies a day.

The cardinal number *five* in both the a- and b-sentence may give rise to scalar inferences consisting in the negation of values higher on the scale. However, as already argued by Horn (1972), the associated scales are different. If the scale associated with the a-sentence is ⟨. . . ,6, 5, 4, . . .⟩, the one related to the b-sentence is the reversal scale ⟨. . . , 4, 5, 6, . . .⟩. In the former case the scalar inference is that Pete does not drink more than five whiskies a day, while in the latter case it can be inferred that Pete did not cut down his drinking to less than five whiskies a day.[28]

Finally, let us consider briefly the problem of scale partitioning. A clear illustration of this phenomenon are natural scales, in particular temperature scales. Temperature scales as a whole do not form a coherent linguistic scale but have to be split up in order to become one. For instance, the temperature scale ⟨. . . , hot, warm, cool, cold, . . .⟩ is not a coherent linguistic scale because it is divided into both warm and cold graduations. The absence of coherence between these values is expressed, for example, by the fact that if something is hot it is also warm but certainly not cold. The criterion of scale coherence is met, when split up in the partial scale ⟨. . . , hot, warm, . . .⟩ and the (reverse ordered) partial scale ⟨. . . , cold, cool, . . .⟩.

In Section 3.3.2 we will give a general coherence criterion for linguistic scales in terms of topics and topic-forming questions. This criterion ac-

[28] The notion of scale reversal is not adopted in Hirschberg (1985), who argues that scale orderings are not defined over their elements but over events which involve them. According to our proposal in Section 3.3.2, this ordering is defined over possible answers to explicit or implicit (sub)topic-forming questions.

counts for the problems mentioned above. However, first we discuss briefly another problem related to the definition of scales, namely that of scale reduction.

3.3.1.2. *Scale Reduction*. Part of the problem of scale definition is the difficult, though hardly discussed problem of scale reduction. It refers directly to the principal question of defining the contextual constraints on the number of values that constitute a linguistic scale. As indicated by others (e.g., Rooth 1992), such constraints are needed in particular to avoid an overgeneration of scalar inferences. For instance, the realization of a value **n** on a scale gives rise to the inference that all (sentences containing) higher values $\mathbf{n} + \mathbf{k}$ ($\mathbf{k} > 0$) on that scale do not obtain. Obviously, this may imply a serious overgeneration of scalar inferences if the set of scale values is a contextually unrestricted one, semantically determined by all possible domain values of the same type.

Central to the argument, therefore, is the problem of defining contextual constraints on the domain entities that in specific cases can function as scalar values. In this respect the notion *focus of attention* and its relation to *discourse structure* is directly relevant. Obviously, entities constituting scale values must be in focus of attention. As demonstrated by Grosz (1978), Reichman (1978), Grosz and Sidner (1986) and others, the set of discourse entities in focus of attention is a variable set the content of which changes in agreement with the structure of discourse. However, what is the relation of this with the notion of topic-forming questions? We have seen how the generation of scalar implicatures is determined by such questions. Though the authors presuppose that a topic is defined for each discourse segment, a formal and operational definition of this notion explicating its relation with that of focus of attention and discourse structure is not given.

In the next section we will characterize a linguistic scale as an ordered topic range bringing into focus of attention an (ordered) set of entities which function as possible extensional values for a questioned un(der)determinacy. Scale reduction which results from a reduction of such a topic range is considered to be a function of the completion task associated with subtopic-forming subquestions. At the highest structural level in discourse, the set of entities in focus of attention is therefore given by the original, contextually unreduced topic range introduced by a main, higher-order, topic-forming question. This range is limited due to subquestions induced at lower levels.

3.3.2. *The Proposal*

Our proposal for linguistic scales that accounts for the view that scalar inferences are contextually determined entailments starts from the following definition which relates the notion of linguistic scales to the notion of topic-forming questions.

Scale Definition. A set of values $\mathbf{S^{T_P}}$ constitutes a linguistic scale iff it meets the following two conditions:

(i) *Condition of Coherence.* The values of $\mathbf{S^{T_P}}$ share the same topic $\mathbf{T_p}$. They constitute possible answers (comments) to the same topic-forming question $\mathbf{Q_p}$. These values represent either satisfactory or less specific, unsatisfactory answers to this question. In the former case the values are identical to those of the actual topic range $\boldsymbol{\rho}'(\mathbf{T_p})$ consisting of 'exact', mutually exclusive answer values. In the latter case they form less specific 'at least' values in terms of which the 'exact' values are grouped.

(ii) *Condition of Ordering.* In contrast to the topic range $\boldsymbol{\rho}'(\mathbf{T_p})$ consisting of 'exact' answer values, the linguistic scale $\mathbf{S^{T_P}}$ is a (partially) ordered set the ordering of which implies that the values from $\rho'(\mathbf{T_p})$ are organized in terms of non-uniquely determining 'at least' values. These values do not possess the property of mutual exclusiveness, i.e. the property that the value selected by a satisfactory answer to $\mathbf{Q_p}$ excludes all other scale values.

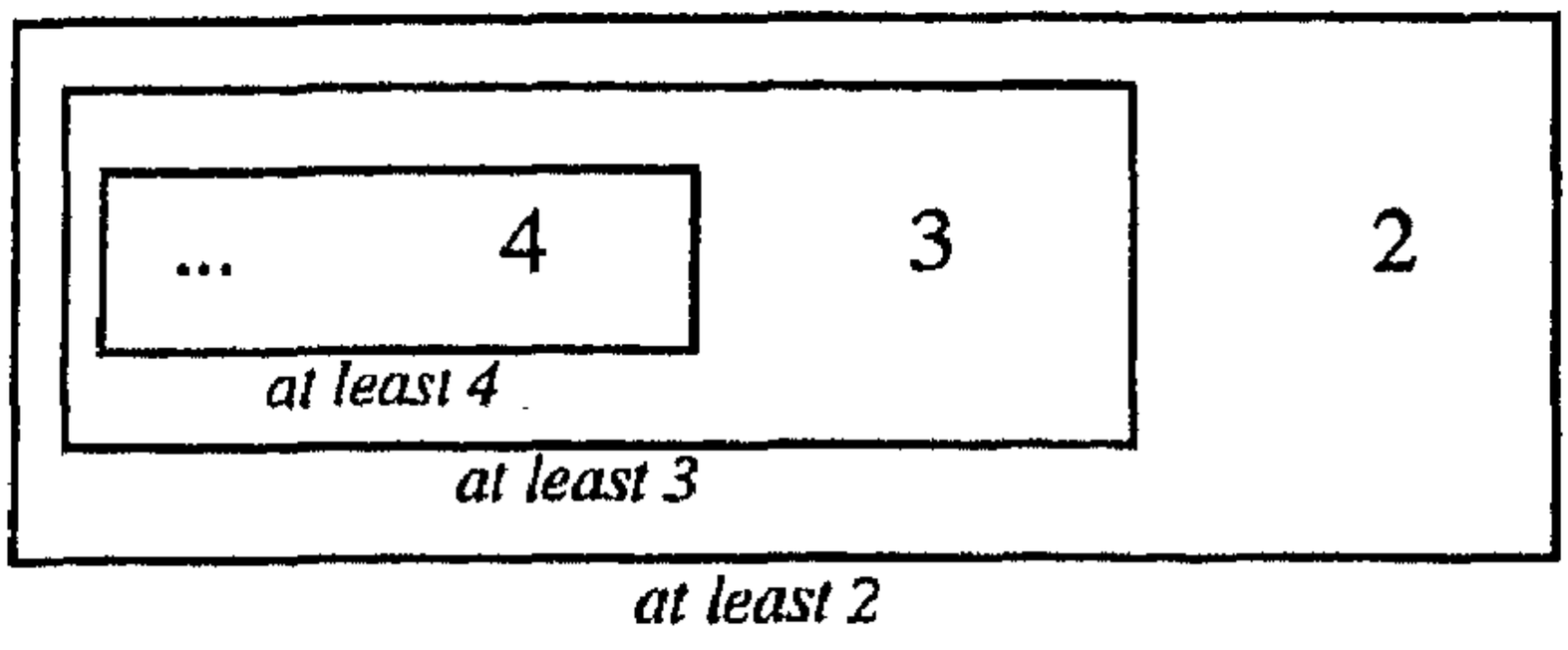
A linguistic scale $\mathbf{S^{T_P}}$ is thus considered to be a (partially) ordered actual topic range $\boldsymbol{\rho}'(\mathbf{T_p})$: $\mathbf{S^{T_P}} = \langle \boldsymbol{\rho}'(\mathbf{T_p}), \geqslant \rangle$. **As will be explained, the scale ordering $\geqslant$ is uniformly defined in terms of** *answer informativeness*, comprising both semantic and pragmatic orderings of the type illustrated above.

According to the definition, the actual topic range $\boldsymbol{\rho}'(\mathbf{T_p})$ consists of 'exact', mutually exclusive answer values. In $\mathbf{S^{T_P}}$ these 'exact' values are ordered in higher ('stronger') and lower ('weaker') values implying an ordering in terms of corresponding, less specific 'at least' values. If, for instance, a scale value '3' is stronger than a scale value '2' this means that '3 is at least 2', implying that the satisfactory answer to question $\mathbf{Q_p}$ is either '2' or '3' (or a higher number). As shown in (41c), such a scale ordering is expressed by inclusion relations between the 'at least' values comprising the corresponding 'exact' values.

(41) a. Q_1 A: How many books does John have to read for the exam, two, three, four or even more books?
A_1 B: John has to read three books for the exam.

b. *Topic range* Q_1:
$\rho'(T_1) = \{2, 3, 4, \ldots\}$

c. *Linguistic scale* Q_1:
$S^{T_1} = \langle\{\ldots, 4, 3, 2\}, \geqslant\rangle$
$= \langle\ldots, 4, 3, 2\rangle$



(41b) shows that the topic range defined by question Q_1 is: $\rho'(T_1) = \{2, 3, 4, \ldots\}$. It is an unordered set containing mutually exclusive ('exact') answer values of which one is selected by the satisfactory answer A_1. If the number of books John has to read for the exam is exactly three, it follows by entailment that this number is neither two nor four, nor more. The corresponding linguistic scale S^{T_1}, on the other hand, is an ordered set which is derived from the topic range $\rho'(T_1)$ and which gives rise to 'at least' values: if the number of books John has to read for the exam is (exactly) four, the number of books he has to read is also at least three and also at least two. As is usual in implicature theories which only consider total orderings, we will represent this scale as an ordered n-tuple in the following way: $S^{T_1} = \langle\ldots, 4, 3, 2\rangle$.

The definition of linguistic scales accounts for the generation of scalar inferences in the following way. In (41) the satisfactory answer A_1 selects the exact comment value '3' on the linguistic scale S^{T_1} introduced by question Q_1. This implies that the number of books John has to read for the exam is at least and at most three. 'At most three' then gives rise to the scalar inference 'not at least four (. . .)', implying the upper bound scalar inference that John does not have to read exactly four or more books. However, answer A_1 also gives rise to an inference with respect to the lower value on S^{T_1}, namely the entailment that if John has to read (exactly) three books for the exam he also has to read (at least) two books. Obviously, no upper bound scalar inference would be induced if the anwer to question Q_1 were a less specific, unsatisfactory answer, e.g. the answer that John does not have to read exactly but *at least* three books.

The scale definition provides a uniform ordering principle for linguistic scales in terms of the general notion of answer informativeness.[29] As demonstrated earlier (Section 2) complex answering processes are determined by a stage-like reduction of the un(der)determinedness of the actual extension of the main topic (expression). This reduction implies a reduction of the associated topic range, realized by subquestions. Answers to subquestions make a preceding unsatisfactory answer more informative, i.e. more specific. This increase of answer informativeness consists in excluding values as possible answers to the main question. The same kind of answer informativeness holds for linguistic scales as a uniform ordering principle.

The general ordering principle in terms of answer informativeness implies that higher scale values are more informative (more specific) than lower values in the sense that they exclude more possible answers as satisfactory answers to the question. First, as is obvious, 'exact' answer values representing a unique value are more informative than lower 'at least' values involving more than one of such a value. Second, higher 'at least' values are more informative than lower 'at least' values. They exclude more possible satisfactory answers to a question than lower 'at least' values. In (41) the non-specific scale values (*at least*) *two*, (*at least*) *three* and (*at least*) *four* give rise to the following sets of possible satisfactory answers to question Q_1: {2, 3, 4, . . .}, {3, 4, . . .} and {4, . . .}. The possible answer *at least four*, e.g., is more specific than *at least three*. The former excludes as possible (satisfactory) answers to the question the values two and three, while the latter excludes only the value two. Clearly, the 'exact', mutually exclusive values on the scale are all equally informative. It holds that every such value excludes all other 'exact' values on the scale.

The definition of linguistic scales contributes, in a direct way, to a solution of the problem of scale activation. We have already talked about a necessary condition for scale activation which is now formulated as follows: given a discourse unit U_i, no other linguistic scale is activated than the one which can be derived from the topic defined for this unit, namely topic T_i determined by the explicit or implicit question Q_i that is answered by U_i. The linguistic scale activated in (41), e.g., is derived from the topic defined by question Q_1, in particular the topic range $\rho'(T_1)$ that is introduced by this question. However, if the topic-forming question Q_1 had been *Who must read three books for the exam, John or both John and*

---

[29] The notion of answer informativeness referred to here differs from Levinson's (1983) notion of scale informativeness which implies scale orderings to be defined by semantic entailment.

his wife? quite a different scale would have been activated, namely $S^{T_1} =$ ⟨{John, John's wife}, {John}⟩. The analysis implies that if the quantifying term *three* in the answer does *not* represent a comment value, no scalar inference is generated on the basis of this term.

The condition for scale activation described here is a necessary but not sufficient one. A comment value may involve a corresponding linguistic scale, though this is not necessarily the case. Another condition for scale activation is provided by the scale definition, namely that the 'exact' values constituting the topic range can be ordered in terms of less specific, 'at least' values. Consider in this respect Hirschberg's (1985) example, to which we have added two different implicit questions, only one of which gives rise to a linguistic scale.

(42) a. ⟨Q_1⟩ A: ⟨Who is the most famous movie star from whom you got an autograph?⟩
Did you get Paul Newman's autograph?
A_1 B: I got Joanne Woodward's.
b. Linguistic scale $S^{T_1} \doteq \langle \ldots, \{PN\}, \{JW\}, \ldots \rangle$

(42)′a. ⟨Q_1⟩ A: ⟨From which movie star did you get an autograph?⟩
Did you get Paul Newman's autograph?
A_1 B: I got Joanne Woodward's.
b. No linguistic scale defined

In contrast to (42), answer A_1 in (42)′ does not give rise to an (upper bound) scalar inference. No linguistic scale is defined here, implying the absence of higher and lower values. Though in both cases it is inferred that B did not get Paul Newman's autograph, the main inferential difference between the two examples is that only the answer in (42) does not exclude lower values, namely that B got an autograph from some movie star lower on the scale, i.e. someone less famous than Joanne Woodward. However, this possibility is absent in case of (42)′. The answer entails that B didn't get an autograph from any other movie star, so also not from someone more or someone less famous than Joanne Woodward.

The definition also provides a solution for the other, earlier mentioned problems related to scale definition, namely the hierarchy problem, the problem of scale overlap, scale direction and scale partitioning. Each of these constitutes a problem of scale coherence implying that the first condition of the definition is not met. In none of the cases the scale values can be interpreted as values sharing the same topic defined by an explicit or implicit topic-forming question.

Unacceptable linguistic scales exhibiting the hierarchy problem, as is

the case with the linguistic scales #⟨Amsterdam, Holland⟩ and #⟨tulips, flowers⟩ given above, consist of values identical to answers to questions which differ from each other in the amount of specificity of the possible answers that can be given. For instance, *Amsterdam* and *Holland* cannot be answers to the same topic-forming question, as is demonstrated by the following unacceptable alternative question: *#Where do you live, in Amsterdam or Holland?* The same holds for the unacceptable linguistic scale #⟨tulips, flowers⟩: *#What did you buy, tulips or flowers?* In other words, in addition to salience as a necessary condition (see above) we add the mutual exclusiveness of the answer values as a sufficient one.

In the case of partial scale overlap the criterion on the basis of which a scale value coheres with the overlapping value in a given context is unclear. This problem was illustrated in (39) in the context of the partially overlapping linguistic scales ⟨warm, lukewarm⟩ and ⟨cold, lukewarm⟩. This problem is now explained as follows. The two overlapping scales are introduced by different topic-forming questions. The coherence condition (i) is not met if the 'wrong' question is associated with the scale inducing sentence. Consider once again example (39), repeated here as example (43).

(43)a. How warm is your beer'?
My beer is lukewarm, if not downright {warm/#cool}
b. How cold is your coffee?
My coffee is lukewarm, if not downright {cold/#hot}

Contrary to example (39) the sentences in (43) are placed in a context of two appropriate, though different, topic-forming questions. Changing the questions would result in incoherent values.

Obviously, the problem of scale direction implying, as demonstrated above, a reversal of scale ordering can also considered to be the effect of different topic-forming questions. The a-sentence in (40), namely *Pete drinks five whiskies a day*, is an appropriate answer to the question *How many whiskies a day does Pete drink?* The b-sentence, on the other hand, answers a different question, e.g. *To how many whiskies a day did Pete cut down his drinking? – Pete cut down his drinking to five whiskies a day.*

Finally, we discussed the coherence problem of scale partitioning occurring, e.g., in relation to temperature scales. The temperature scale ⟨. . . , hot, warm, cool, cold, . . .⟩ is not a coherent linguistic one because its values cannot be interpreted as possible answers given to the same topic-forming question. Answers corresponding to elements on the partial scale ⟨. . . , hot, warm, . . .⟩ answer a '*How warm*'-question, while those

corresponding to the partial scale ⟨. . . , cold, cool, . . .⟩ answer a reverse '*How cold*'-question.

In addition to the problem of scale ordering, scale activation and scale coherence, we also pointed out the important issue of scale reduction by contextual factors. Our criterion for contextual constraints on linguistic scales is given by the completion task associated with subtopic-forming subquestions. As argued above, these questions imply a stage-like reduction of the topic range associated with a higher-order topic-forming question. As a consequence, they realize a reduction of the linguistic scale derived from this range.[30] Example (44) gives a simple illustration of the contextual reduction of a linguistic scale in a process of subquestioning.

(44)

Q_1	A: How much can you pay? (\$85, \$90, \$95, \$100, \$105, \$110 or more)	$S^{T_1} = \langle . . . , \$110, \$105, \$100, \$95, \$90, \$85\rangle$
A_1	B: Less than \$100.	$S^{T_1'} = \langle \$95, \$90, \$85\rangle$
Q_2	A: \$95?	
A_2	B: No, that's too much.	$S^{T_1''} = \langle \$90, \$85\rangle$
Q_3	A: How much then?	
A_3	B: \$90.	

Reduction of the linguistic scale S^{T_1} corresponds to a reduction of the corresponding topic range $\rho'(T_1)$.

3.3.3. *Partially Ordered Topic Ranges as Linguistic Scales*

So far, we have only considered linguistic scales as linear ordered sets which are conventionally represented as *n*-tuples. However, in practice partially ordered sets are more common than these total orderings. The former contain both related and unrelated, so-called comparable elements. As pointed out by Hirschberg (1985), we need to broaden the notion of scalar inferences to arrive at one which also accounts for inferences based on these partial orderings.[31] Our definition of linguistic scales given above accounts for inferences of this type in terms of partially ordered topic

---

[30] Obviously, this reduction only concerns higher scale values, because the lower ones represent at least values entailed by those selected by the (un)satisfactory answer.

[31] Despite the fact that early implicature theory concentrates on total orderings, the phenomenon of scales as partial orderings has not been unobserved. For instance, Gazdar (1979: 58, n. 22): "Note that treating scales as *n*-tuples obscures the fact that, in general, we are dealing with partial rather than total orderings".

ranges introduced on different discourse levels by an explicit or implicit topic-forming question.

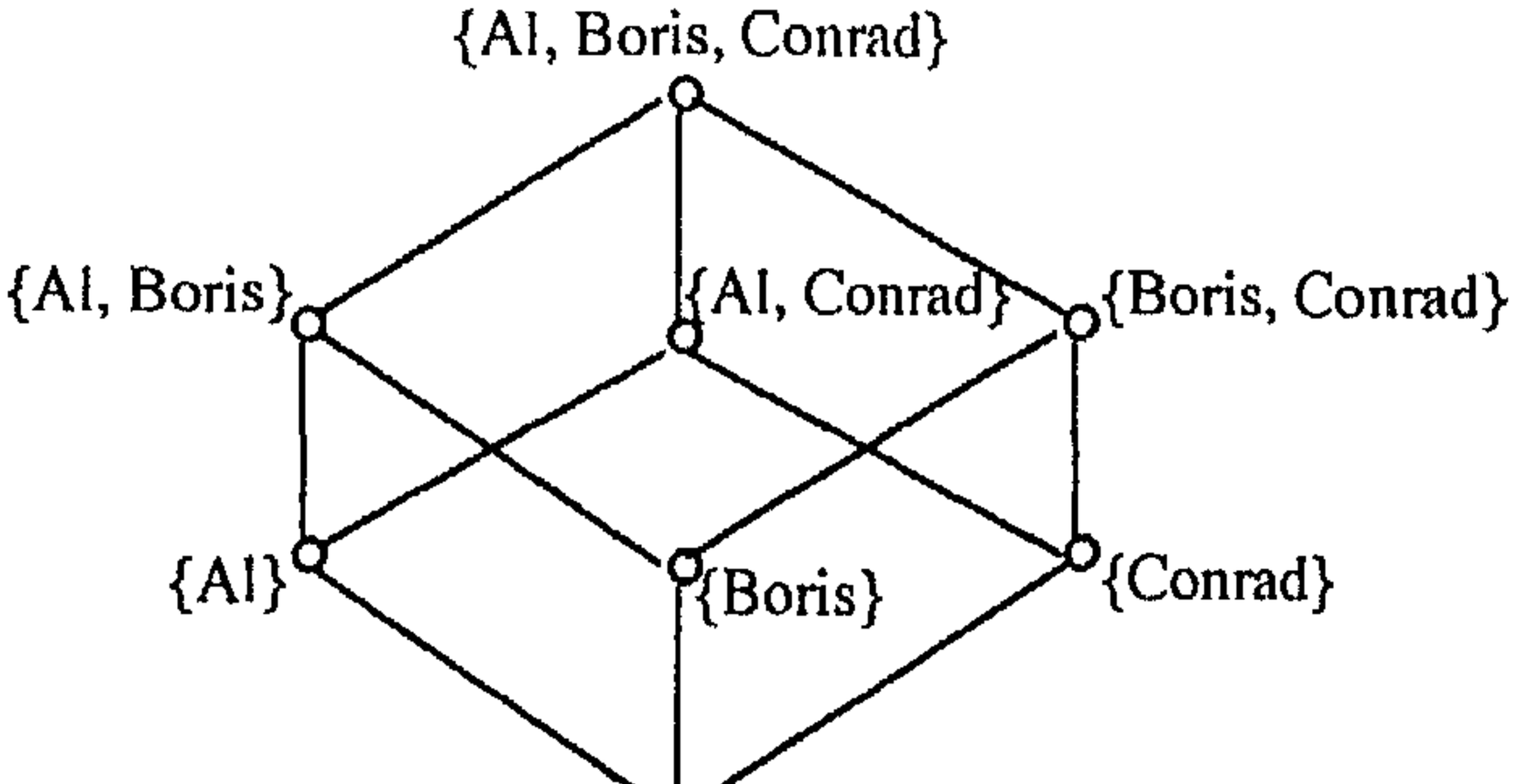
As a brief illustration consider the question Q_r:*Who complained?* asked in a context which is unrestricted by discourse structure and in which only the set of persons $A = \{\text{Al, Boris, Conrad}\}$ is relevant. Assuming that someone did complain, the corresponding actual topic range is defined as follows: $\rho'(T_r) = \{X \mid X \subseteq A \wedge |X| \geqslant 1\}$. The related linguistic scale is a partially ordered set derived from this topic range: $S^{T_r} = \langle \rho'(T_r), \geqslant \rangle$. It does not meet the condition of linearity characteristic for total orderings, namely that for all $x, y \in A$: $x \geqslant y$ or $y \geqslant x$. In (45) the partially ordered linguistic scale S^{T_r} is visually represented by a Hasse diagram.[32]

(45)a. Q_r: Who complained?

Linguistic scale Q_r:

$$S^{T_r} = \langle \{X \mid X \subseteq \{\text{Al, Boris, Conrad}\} \wedge |A| \geqslant 1\}, \geqslant \rangle$$

Hasse diagram of S^{T_r}:



The diagram shows the structure of ordered sets which is of a hierarchical nature.

As in the case of total orderings, the partial ordering implies that the 'exact' values on the scale are organized in terms of 'at least' values: higher 'exact' values entail lower 'at least' values. For instance, if answer A_r to question Q_r is that both Al and Boris complained, it follows by entailment that at least Al and at least Boris complained. However, answer A_r also gives rise to an inference consisting of the negation of the sentence which realizes a value higher on the scale, namely that all three persons complained. This upper bound scalar inference implies that it was not both Al and Conrad who complained, not both Boris and Conrad, nor

[32] See, in particular, Partee, ter Meulen and Wall (1990) on this point.

extended answer. Second, contrary to what might be expected, this does not imply that the moment of induction takes place directly after the production of this discourse part. We present two inference rules needed for adequately computing a global scalar inference.

In Section 3.4.2, on the other hand, brief attention will be paid to the phenomenon of local scalar inferences discussed above, in particular to their behavior in hierarchical topic processes. We will discuss local scalar inferences resulting from both quantitatively and qualitatively unsatisfactory answers, looking, in particular, at their relation to global scalar inferences.

3.4.1. *Determination of Global Inferences: The Inducing Context and the Moment of Induction*

Global scalar inferences are determined either by a part of the extended answer or the answer as a whole. We first consider the situation in which a global scalar inference is generated on the basis of the whole answer. In such a situation the extended answer consists of a *quantitatively* unsatisfactory answer providing an incomplete comment value and an elaboration on it providing the remaining value(s). Together, these partial comment values constitute the requested final comment value on the basis of which a global scalar inference is generated. The generation of a global scalar inference as the result of the whole answer occurs in our example (4)′ given above which is repeated here as example (46).

(46)
F_1 A: A well-known subsidy book publisher is searching for manuscripts.
Q_1 B: What kind of manuscripts?
A_1 A: Fiction and non-fiction will be considered.
Q_2 B: What else?
A_2 A: Poetry, juvenile, travel, scientific, specialized and even controversial subjects.

The insufficient answer *Fiction and non-fiction* together with the completing answer *Poetry, juvenile, travel, scientific, specialized and even controversial subjects* provide the final comment value that determines the higher-order upper bound scalar inference associated with this discourse unit. This upper bound scalar inference consists of the negation of sentences containing a scalar value higher on the linguistic scale introduced by the main question Q_1, namely those comprising a superset of the set of manuscripts referred to by the extended answer.

But, now consider again our earlier example (4)″, here example (47).

(47) F_1 A: A well-known subsidy book publisher is searching for potentially successful manuscripts.
Q_1 B: What kind of manuscripts?
A_1 A: Only fiction.
Q_2 B: What kind of fiction?
(I heard that the success of some types of fiction has been decreasing in recent months.)
A_2 A: Both novels and short stories.

In contrast to the preceding example, (47) involves an extension of a *qualitatively* unsatisfactory answer. In this case the upper bound scalar inference associated with the extended answer to the main question Q_1 is determined by only a *part* of this answer, namely the answer *Both novels and short stories* (*will be considered*) given eventually by answer A_2. It entails the less specific unsatisfactory answer *Fiction will be considered* which may be deleted without loss of inferential force on the global level. It provides the final comment value to the topic introduced by question Q_1 and, as a consequence, gives rise to the upper bound scalar inference on the global level.

In (47) the inference determining part of the answer comes at the end, though this is not always the case in answering processes involving a qualitative extension of an unsatisfactory answer. As argued in Van Kuppevelt (1994), an answer may be qualitatively unsatisfactory not because it is insufficiently specific, as is illustrated in (47), but because it has not yet been fully accepted by the addressee and therefore calls for support, e.g. a justification. In the latter case the part of the answer that finally determines the global scalar inference is not located at the end of the discourse but in the beginning. Consider the related example (48).

(48) F_1 A: A well-known subsidy book publisher is searching for manuscripts.
Q_1 B: What kind of manuscripts?
A_1 A: Fiction will be considered.
Q_2 B: Are you sure about this?
A_2 A: Yes, ⟨Why?⟩ someone told me.

Answer A_1 is unsatisfactory as appears from subquestion Q_2. It is not yet fully accepted by questioner B and calls for a justification, which is given in A_2. In this case not A_1 but A_2 may be deleted without loss of inferential force on the global level. Answer A_1 provides the final comment value

determining the scalar inference associated with the extended answer to the main question Q_1.[34]

So far, our conclusion is that the upper bound scalar inference associated with an extended answer to a higher-order topic-forming question is determined either by the complete answer, in the case of an quantitatively unsatisfactory answer, or by just a part of it in the case of the answer being qualitatively unsatisfactory. As illustrated above, this part is presented at the beginning or the end of the answering process. The following inference rule concerning the computation of scalar inferences at the global level provides the generalization that is needed.

Inference Rule 1: When computing a global scalar inference associated with a discourse unit answering an explicit or implicit higher-order question, the computational input which actually determines this inference is the *final comment value* to the topic defined by this question.

In (46) the final comment value is provided by the complete answer to the main question, while in (47) and (48) it is given in, respectively, the first part (A_1) and last part (A_2) of the extended answer to these questions. These parts, in particular, cannot be deleted without a loss of inferential force, since they satisfy that which is asked for by the main question.

Although a global scalar inference is determined by the part of the extended answer that provides the final comment value, the moment that such an inference can be determined is always just after the whole answer to the question has been given. Premature determinations may give rise to false predictions. For instance, if in (46) the global scalar inference were to be determined while considering the quantitatively unsatisfactory answer A_1 as being the one providing the final comment value, a contradiction would arise. The prediction would be an upper bound scalar inference contradicting the answer given to subquestion Q_2, namely the inference that the publisher in question is not also searching for manuscripts of *Poetry*, *etc.*

More interesting is the premature determination of a global scalar infer-

---

[34] As accounted for extensively in Van Kuppevelt (1994), the phenomenon illustrated in (46), (47) and (48) is that of *directionality*. The phenomenon is called *nuclearity* in Rhetorical Structure Theory (e.g., Mann and Thompson 1988), and, as explicated in Moser and Moore (1993), is directly related to the notion of *dominance* central to the intentional approaches of discourse structure (see, in particular, Grosz and Sidner 1986). It refers to the property of (a part of) a text to be directed toward a goal, mostly resulting in asymmetric functional relations between related discourse segments. We can distinguish three types of directionality occurring in (46), (47) and (48), namely *bi-directionality*, *forward directionality* and *backward directionality*.

determining the scalar inference associated with the extended answer to the main question Q_1.[34]

So far, our conclusion is that the upper bound scalar inference associated with an extended answer to a higher-order topic-forming question is determined either by the complete answer, in the case of an quantitatively unsatisfactory answer, or by just a part of it in the case of the answer being qualitatively unsatisfactory. As illustrated above, this part is presented at the beginning or the end of the answering process. The following inference rule concerning the computation of scalar inferences at the global level provides the generalization that is needed.

Inference Rule 1: When computing a global scalar inference associated with a discourse unit answering an explicit or implicit higher-order question, the computational input which actually determines this inference is the *final comment value* to the topic defined by this question.

In (46) the final comment value is provided by the complete answer to the main question, while in (47) and (48) it is given in, respectively, the first part (A_1) and last part (A_2) of the extended answer to these questions. These parts, in particular, cannot be deleted without a loss of inferential force, since they satisfy that which is asked for by the main question.

Although a global scalar inference is determined by the part of the extended answer that provides the final comment value, the moment that such an inference can be determined is always just after the whole answer to the question has been given. Premature determinations may give rise to false predictions. For instance, if in (46) the global scalar inference were to be determined while considering the quantitatively unsatisfactory answer A_1 as being the one providing the final comment value, a contradiction would arise. The prediction would be an upper bound scalar inference contradicting the answer given to subquestion Q_2, namely the inference that the publisher in question is not also searching for manuscripts of *Poetry*, *etc.*

More interesting is the premature determination of a global scalar infer-

---

[34] As accounted for extensively in Van Kuppevelt (1994), the phenomenon illustrated in (46), (47) and (48) is that of *directionality*. The phenomenon is called *nuclearity* in Rhetorical Structure Theory (e.g., Mann and Thompson 1988), and, as explicated in Moser and Moore (1993), is directly related to the notion of *dominance* central to the intentional approaches of discourse structure (see, in particular, Grosz and Sidner 1986). It refers to the property of (a part of) a text to be directed toward a goal, mostly resulting in asymmetric functional relations between related discourse segments. We can distinguish three types of directionality occurring in (46), (47) and (48), namely *bi-directionality*, *forward directionality* and *backward directionality*.

ence in case of (48). Answer A_1 in fact provides the final comment value to the topic introduced by the main question Q_1. However, the global inference must be computed at the end, namely after answer A_2 has been given. The justification this answer provides may be denied by the addressee and, as a consequence, would rule out that A_1 provides the final comment value.

The conclusion, therefore, is that in determining global scalar inferences, the inference rule relating to the inducing context given above does not suffice. In addition to this rule we need one relating to the moment of induction, namely the following:

Inference Rule 2: When computing a global scalar inference, the moment of computation must be just after the whole answer has been produced. This moment does not necessarily coincide with the moment of realization of the final comment value.

Thus inference Rule 2 implies that scalar inferences must be computed after segment closure which coincides with the moment that the full answer to the main question has been realized.

3.4.2. *Determination of Local Inferences in Hierarchical Contexts*

In the preceding section we focused on the determination of global scalar inferences, characterized as inferences induced as the result of extended answers given to higher-order topic-forming questions. We argued that the computation of such inferences necessarily requires the identification of that part of the answer which provides the final comment value to the topic defined by the higher-order question. As illustrated, this value is not necessarily provided by the whole extended answer given to this higher-order question. The global scalar inference is computed by determining the inferential output of this value with respect to the linguistic scale activated by the higher-order question.

In this last section we briefly discuss the determination of local scalar inferences which are generated on-line in the discourse process and which directly contribute to a scalar inference on the global level. Obviously, the inferential result of locally induced scalar inferences is identical to a global one associated with an extended answer. As compared to global inferences, this type of inference results from non-extended answers, e.g. incomplete and non-specific answers as well as individual extensions given to such answers. They result from the comment values provided by these

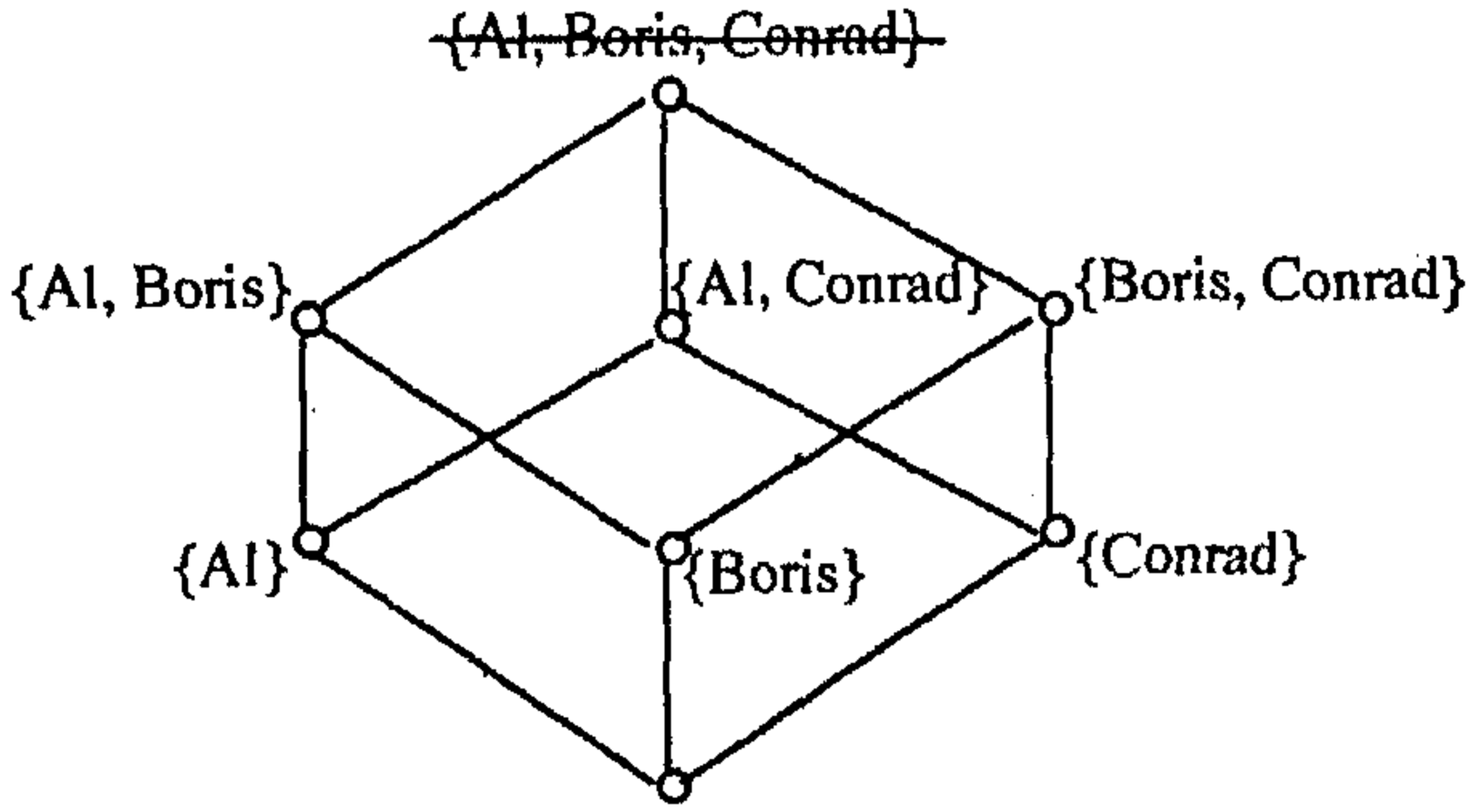
answers and are computed with respect to the linguistic scale associated with the local question answered by the non-extended answer. Because inferences of this type were extensively discussed in previous sections, we will now pay attention only to how they relate to global scalar inferences.

Let us consider in this respect two typical variants of our preceding example in which an answer is given to the question *Who complained?*. It is assumed again that this question is asked in the context in which only the set of persons $A = \{\text{Al, Boris, Conrad}\}$ is relevant. The first variant involves a qualitative extension of an unsatisfactory answer to this question, namely one which contributes to the non-specific answer.

(49)a. . . .
Q_1 B: Who complained?
A_1 A: Two persons.
Q_2 B: Which two persons?
A_2 A: Al and Boris.

The unsatisfactory answer A_1 reduces the topic range $\rho'(T_1)$ to the set of possible answers {{Al, Boris}, {Boris, Conrad}, {Al, Conrad}}. The local scalar inference induced on the basis of this answer consists only of the negation of (the sentence containing) the highest value {Al, Boris, Conrad} on the corresponding scale S^{T_1}.

b.



Together with the local scalar inferences induced as the result of the answer given to subquestion Q_2 this inference amounts to an inferentional result that is identical to that implied by the scalar inference on the global level. This was presented in (45b).

The second variant consists of the completion of a quantitatively unsatisfactory, partial answer.

(49)'a.

. . .

Q_1 B: Who complained?
A_1 A: Al in particular.
Q_2 B: Who else?
A_2 A: Boris

In this case, the unsatisfactory answer A_1 reduces the original range of topic T_1 to the actual topic range $\rho'(T_1)$ = {{Al, Conrad}, {Al, Boris}, {Al, Boris, Conrad}}. However, no local (upper bound) scalar inference is generated as the result of this incomplete answer. The linguistic scale S^{T_1} which is derived from $\rho'(T_1)$ still contains the strongest possible value {Al, Boris, Conrad} implying that no lower ('at least') value on the scale is excluded at this stage in the development of the discourse. Answer A_2, on the other hand, gives rise to an inferential result which is identical to the one implied by the upper bound scalar inference on the global level.

4. Conclusion

In this paper we proposed an alternative definition of linguistic scales accounting for the generation of what is generally known as scalar implicatures. The proposal was stated within a topical approach of discourse structure explaining the structural coherence in discourse in terms of topic-forming explicit or implicit questions. According to the central hypothesis of this approach, the topic of a discourse unit is provided by the topic-forming explicit or implicit question answered by that unit, while the relation between (hierarchically organized) discourse units is determined by the relation between these topic-forming questions. It has been demonstrated that by relating the notion of linguistic scales to that of topic-forming questions it is possible to compute the scalar inferences associated with different discourse levels, thereby going beyond an account of locally induced scalar inferences which is standard in pragmatic theory.

We have presented evidence for the view that scalar implicatures, both those induced as the result of cardinals and those induced as the result of non-cardinals, are in fact determined in a uniform way by topic-forming questions. We discussed the important issue of overgeneration of scalar inferences as well as that of the inferential status of inferences of this type. As far as overgeneration is concerned, we claimed that so-called upper bound scalar inferences are not induced under all topic-comment modulations of the inference inducing question-answering sentence, but only as the result of what forms the comment part of this sentence in a

given context. However, a given comment value may be unsatisfactory and a question may be functionally subservient to a higher-order one involving a weakening of the topic introduced by it. We illustrated that in both cases an upper bound scalar inference may be blocked. As to the inferential status of scalar inferences, we have argued that they are semantic entailments rather than weaker pragmatic inferences which in principal can be cancelled without giving rise to a contradiction. In this context we have discussed both Horn's (1972) 'elimination by contradiction' and 'elimination by suspension' cases and Gazdar's arguments against a purely semantic treatment of these inferences.

Scalar inferences associated with higher- or lower-order discourse units are generated on the basis of linguistic scales introduced by topic-forming questions defining these units. A linguistic scale is defined as a (partially) ordered topic range. This definition made it possible to provide adequate solutions to problems related to scale definition, including the problems of scale activation, scale reduction, scale ordering and scale coherence. Furthermore, the definition gave rise to an explanation of the way in which upper bound scalar inferences induced at the local level contribute to global ones.

References

Atlas, J. D.: 1990, 'Implicature and Logical Form: The Semantics-Pragmatics Interface', Lectures presented at the Second European Summer School in Language, Logic and Information, Katholieke Universiteit Leuven.

Bartsch, R.: 1976, 'Topik-Fokus-Struktur und kategoriale Syntax', in V. Ehrich and P. Finke (eds.), *Grammatik und Pragmatik*, Scriptor Verlag, Kronberg.

Belnap, N. D. and T. B. Steel: 1976, *The Logic of Questions and Answers*, Yale University Press, New Haven.

Belnap, N. D.: 1982, 'Questions and Answers in Montague Grammar', in S. Peters and E. Saarinen (eds.), *Processes, Beliefs, and Questions*, Reidel, Dordrecht.

Campbell, R. N.: 1981, 'Language Acquisition, Psychological Dualism and the Definition of Pragmatics', in H. Parret, M. Sbisà, and J. Verschueren (eds.), *Possibilities and Limitations of Pragmatics*, John Benjamins, Amsterdam.

Carston, R.: 1985, 'A reanalysis of Some "Quantity Implicatures"', Unpublished ms., University College, London.

Carston, R.: 1988, 'Implicature, Explicature, and Truth-Theoretic Semantics', in R. M. Kempson (ed.), *Mental Representations: The Interface between Language and Reality*, Cambridge University Press, Cambridge.

Fauconnier, G.: 1975a, 'Pragmatic Scales and Logical Structure', *Linguistic Inquiry* **6**, 353–375.

Fauconnier, G.: 1975b, 'Polarity and the Scale Principle', *Papers from the 11th Regional Meeting, Chicago Linguistic Society*, 188–199.

Fretheim, T.: 1992, 'The Effect of Intonation on a Type of Scalar Implicature', *Journal of Pragmatics* **18**, 1–30.

Gazdar, G.: 1979, *Pragmatics: Implicature, Presupposition and Logical Form*, Academic Press, London.
Grice, H. P.: 1967, *Logic and Conversation*, Unpublished ms. of the William James Lectures, Harvard University.
Grice, H. P.: 1975, 'Logic and Conversation', in P. Cole and J. L. Morgan (eds.), *Syntax and Semantics 3: Speech Acts*, Academic Press, New York.
Groenendijk, J. A. G. and M. J. B. Stokhof: 1984, 'On the Semantics of Questions and the Pragmatics of Answers', *Studies on the Semantics of Questions and the Pragmatics of Answers*, Ph.D. Thesis, University of Amsterdam, pp. 209–250.
Grosz, B. J. and C. L. Sidner: 1986, 'Attention, Intentions, and the Structure of Discourse', *Computational Linguistics* **12**, 175–204.
Grosz, B.: 1978, 'Discourse Knowledge', in D. E. Walker (ed.), *Understanding Spoken Language*, Elsevier North-Holland, New York.
Hamblin, C. L,.: 1973, 'Questions in Montague English', *Foundations of Language* **10**, 41–53.
Harnish, R. M.: 1979, 'Logical Form and Implicature', in K. Bach and R. M. Harnish (eds.), *Linguistic Communication and Speech Acts*, MIT Press, Cambridge (Mass.).
Hausser, R. R.: 1983, 'Surface Compositionality and the Semantics of Mood', in J. R. Searle, F. Kiefer, and M. Bierwisch (eds.), *Speech Act Theory and Pragmatics*, Reidel, Dordrecht.
Hirschberg, J.: 1985, *A Theory of Scalar Implicature*, Ph.D. Thesis, University of Pennsylvania. (Published in the Outstanding Dissertations in Linguistics series, Garland Publishers, New York, 1991).
Horn, L. R.: 1972, *On the Semantic Properties of Logical Operators in English*, Ph.D. Thesis, University of California at Los Angeles.
Horn, L. R.: 1989, *A Natural History of Negation*, The University of Chicago Press, Chicago.
Horn, L. R.: 1992, 'The Said and the Unsaid', in C. Barker and D. Dowty (eds.), *Proceedings of SALT II*, Ohio State University.
Karttunen, L.: 1977, 'Syntax and Semantics of Questions', *Linguistics and Philosophy* **1**, 3–44.
Kempson, R.: 1986, 'Ambiguity and the Semantics-Pragmatics Distinction', in C. Travis (ed.), *Meaning and Interpretation*, Blackwell, Oxford.
Klein, W. and C. von Stutterheim: 1987, 'Quaestio und referentielle Bewegung in Erzählungen', *Linguistische Berichte* **109**, 163–183.
Kroch, A.: 1972, 'Lexical and Inferred Meanings for Some Time Adverbs', *Quarterly Progress Report of the Research Lab. of Electronics*, MIT, 104.
Levinson, S.C.: 1983, *Pragmatics*, Cambridge University Press, Cambridge.
Mann, W. C. and S. A. Thompson: 1988, 'Rhetorical Structure Theory: Toward a Functional Theory of Text Organization', *Text* **8**, 243–281.
Moser, M. and J. Moore: 1993, 'Investigating Discourse Relations', in O. Rambow (ed.), *Intentionality and Structure in Discourse Relations*, Proceedings of the ACL SIG Workshop, Columbus (Ohio).
Partee, B. H., A. ter Meulen, and R. F. Wall: 1990, *Mathematical Methods in Linguistics*, Kluwer, Dordrecht.
Reichman, R.: 1978, 'Conversational Coherency', *Cognitive Science* **2**.
Rooth, M.: 1992, 'A Theory of Focus Interpretation', *Natural Language Semantics* **1**, 75–116.
Sadock, J. M.: 1978, 'On Testing for Conversational Implicature', in P. Cole (ed.), *Syntax and Semantics 9: Pragmatics*.
Sadock, J. M.: 1984, 'Whither Radical Pragmatics?', in D. Schiffrin (ed.), *Meaning, Form, and Use in Context: Linguistic Applications*, Georgetown University Press, Washington.
Scha, R. J. H.: 1983, *Logical Foundations of Question Answering*, Ph.D. Thesis, University of Groningen.
Scharten, R.: forthcoming, *A Discourse-Semantic Account of Gricean Implicatures*, Ph.D. Thesis, University of Nijmegen.

Seuren, P. A. M.: 1993, 'Why Does 2 Mean "2"? – Grist to the Anti-Grice Mill', in E. Hajičová (ed.), *Functional Description of Language*, Charles University, Prague.
Sperber, D. and D. Wilson: 1986, *Relevance. Communication and Cognition*, Blackwell, Oxford.
Stout, G. F.: 1896, *Analytic Psychology*, Allen and Unwin, London/Macmillan, New York.
Tichý, P.: 1978, 'Questions, Answers and Logic', *American Philosophical Quarterly* **15**, 275–284.
Van Kuppevelt, J.: 1991, *Topic en Comment: Expliciete en Impliciete Vraagstelling in Discourse*, Ph.D. Thesis, University of Nijmegen.
Van Kuppevelt, J.: 1993, 'About a Uniform Conception of S- and D-Topics', in E. Hajičová (ed.), *Functional Description of Language*, Charles University, Prague.
Van Kuppevelt, J.: 1993, 'Topic and Comment', in R. E. Asher (ed.), *The Encyclopedia of Language and Linguistics*, Pergamon Press, Oxford.
Van Kuppevelt, J.: 1994, 'Directionality in Discourse', in P. Bosch and R. van der Sandt (eds.), *Proceedings of the Interdisciplinary Conference on Focus and Natural Language Processing*, Heidelberg: IBM Working Papers.
Van Kuppevelt, J.: 1995a, 'Discourse Structure, Topicality and Questioning', *Journal of Linguistics* **31**, 109–147.
Van Kuppevelt, J.: 1995b, 'Main Structure and Side Structure in Discourse', *Linguistics* **33**, 809–833.
Van Kuppevelt, J.: forthcoming, 'Un(der)determinedness and Questioning in Discourse'.
Vennemann, T.: 1975, 'Topics, Sentence Accent, Ellipsis: A Proposal for their Formal Treatment', in E. L. Keenan (ed.), *Formal Semantics of Natural Language*, Cambridge University Press, Cambridge.

University of Nymegen
Department of Philosophy
P.O. Box 9103
6500 HD Nymegen
The Netherlands
jvkuppev@vms.uci.kun.nl