On Another Pragmatic Facet of Scalar Inference

Dae-Young Kim
(Jeonju University)

Kim, Dae-Young. (2016). On Another Pragmatic Facet of Scalar Inference. *The Linguistic Association of Korea Journal, 24*(4), 151-178. The types of scalar inference can be divided into these two: scalar entailment and scalar implicature. According to Gazdar (1979), Levinson (1983, 2000) and Horn (1985, 1989, 2004), assuming a scale $<e_1, e_2, e_3 \cdots e_n>$, where e_1 scalar-entails e_2, e_2 also scalar-entails e_3, *etc*, but not vice versa. On the other hand, uttering a sentence including (e_n) scalar-implicates the negation of a sentence including (e_{n-1}). These scalar inferences can be the evidence verifying that ordinary language users observe the Maxim of Quantity proposed by Grice (1975), and these pragmatic principles seem to always enable the ordinary language users to foresee regular conclusions, in which any scalar utterances occur. In our ordinary language use, however, sometimes there might be some exceptional cases where the hearer cannot properly interpret the speaker's exact intention, if these scalar inferential principles are to be mechanically applied. The reason is due to the point that various non-linguistic factors such as language users' intuition and their socio-cultural environments can also be involved in the process of the interpreting the scalar utterances, besides the linguistic principles proposed by Horn (1989, 2004) and Levinson (2000). These non-linguistic factors are very significant in that they might influence the ultimate meaning intended by the speaker; in this paper, I discuss what they are, and account for how the scalar inference connected to them can be treated in the discourse.

Key Words: scalar entailment, scalar implicature, metalinguistic negation, Maxim of Quantity, Principle of Politeness, Horn's scale, Matsumoto's scale, pragmatic constraint by the same criterion

1. Introduction

This paper aims to survey the pragmatic nature of scalar inference (including scalar entailment and scalar implicature) which has been developed by Gazdar (1979), Levinson (1983, 2000) and Horn (1985, 1989, 2004), and to point out that there is another pragmatic facet in scalar inference that should be noted, apart from the pragmatic principles of scalar inference which they have discussed.

Before setting to the full-dress discussion, I take one example with reference to scalar inference. Assuming that somebody says "Tom broke three windows", this utterance entails 'Tom broke two windows', and implicates 'Tom did not break more than three windows' at the same time. According to Grice (1975)'s Maxim of Quantity, the hearer who interprets the utterance believes that the rational speaker must have said the necessary and enough content which is as informative as is required for the current purposes of the exchange. Thus, if 'Tom broke three windows' is true, the speaker should say "Tom broke three windows" in order that the speaker may observe the Maxim of Quantity. Furthermore, as long as the hearer knows that the speaker observes Cooperative Principle in the conversation, when the speaker says "Tom broke three windows", the hearer may realize that this utterance implicates 'Tom did not broke four or more than four windows', considering the Maxim of Quantity.

Scalar inference connected to the Maxim of Quantity has been regarded as one of the most crucial notions in neo-Gricean pragmatics which succeeds to the Gricean pragmatic theory, and includes scalar entailment and scalar implicature; from this point of view, 'scale' is a very crucial and useful notion for explaining pragmatic inference. The definition of Horn (1989)'s 'scale', which is most referred, is as follows:

> Quantitative scales are defined by entailment; P_j outranks P_i on a given scale iff a statement containing an instance of the former unilaterally entails the corresponding statement containing the latter. (Horn 1989: 231)

In addition, Levinson (1983) defines the notion of 'scale' as follows:

> A linguistic scale consists of a set of linguistic alternates, or contrasive

> expressions of the same grammatical category, which can be arranged in a linear order by degree of informativeness or semantic strength. Such a scale will have the general form of an ordered set (indicated by angled brackets) of linguistic expressions or scalar predicates, e_1, e_2, e_3 ... e_n, as in: <e_1, e_2, e_3 ... e_n> (Levinson 1983: 133)

Furthermore, these examples given below are some representative scales shown in English:

(1) <all, most, many, some, few>
<and, or>
<n, ... 5, 4, 3, 2, 1>
<excellent, good>
<hot, warm>
<always, often, sometimes>
<succeed in *V*ing, try to *V*, want to *V*>
<necessarily P, P, possibly P>
<certain that P, probable that P, possible that P>
<must, should, may>
<cold, cool>
<love, like>
<none, not all> (Levinson 1983: 134)

According to Gazdar (1979) and Horn (1985, 1989), a semantically strong expression P and Q whose semantic strength is relatively weaker than that of P may form a scale <P, Q>, and this scale brings about scalar inferences: scalar entailment and scalar implicature. When the speaker chooses the stronger expression P, not Q, this utterance scalar-entails that 'if P is true, then Q is inevitably also true'. On the other hand, if the speaker says weaker expression Q, instead of P, this utterance may scalar-implicate that the speaker does not mean P. Typical examples of scalar inference are given below:

(2) Scale: <all, many, some>
a. Many students passed the exam.

b. Some students passed the exam.
c. Not all student passed the exam.

(3) Scale: <five, four, three>
a. John ate four apples.
b. John ate three apples.
c. John did not eat five or more than five apples.

Each (a) in (2) and (3) scalar-entails each (b), and at the same time scalar-implicates each (c).

Furthermore, each (c) as the cases of scalar implicature can be viewed as a type of conversational implicature; namely, they are distinguished from logical entailment in that they are cancelable, and from conventional implicature because they are not detachable. For instance, if I say "Jerry has two children", my utterance logically entails that 'Jerry has one child', and scalar implicates that 'Jerry does not three or more than three children'. However, that scalar implicature can be cancelled in this way: "Jerry has two children; in fact, however, he has three because his wife is pregnant now".

In relation to non-detachability, assuming a scale of <all, most, many, some>, all the following examples regularly scalar-implicate the same meaning, although they are respectively described in different expressions:

(4) a. Some consumers like the new product of X company.
b. Many consumers like the new product of X company.
c. Most consumers like the new product of X company.
↦ 'Not all consumers like the new product of X company'.[1)]

The problem is, however, some cases of scalar implicature cannot fully reflect ordinary language users' intuition. As pointed out in Lee (2001: 235), for example, assuming a scale of <scorching, hot, warm, chilly, cold, freezing>, if the speaker says "It's hot today", this utterance should scalar-entail 'It's cold/freezing today' as well as 'It's warm today' at the same time, in accordance with the principle of scalar inference; but this is definitely far from ordinary

1) This symbol ↦ means 'X conversationally implicates Y'.

language users' intuition. Moreover, it is not the case that "It's freezing today" may scalar-implicate 'It's not hot today'.

In addition to the explanatory problem of scalar inference discussed above, there exists another one; namely, there are some exceptional cases in which the principle of scalar inference by Horn's approach does not successfully work, particularly when a scale is formed by a pragmatic criterion, not by a semantic one. For instance, assuming a scale <Baker Street 221B, London, England>, when Jerry says "Spike studies pragmatics in London", this utterance cannot scalar-implicate 'not in Baker Street 221B'. In this case, it is necessary to pursue another way of explaining scalar inference, because this scale is formed not by the difference of semantic strength but by a pragmatic criterion: the ordinary language users' knowledge about the world. For this reason, it should be noted that sometimes the regular mechanism of scalar implicature could bring about some interpretive problems in the process of our communication, and the principle of scalar inference by some neo-Gricean pragmaticists such as Horn (1989, 2004) and Levinson (1983, 2000) needs reconsidering.

In this paper, with reference to the issue of scalar inference, I claim that some cases of scalar inference should be restricted or reconsidered, in accordance with the given contextual information, instead of merely applying the fixed principle of scalar inference to the given case, because sometimes mere application of the principle cannot guarantee exactly interpreting the speaker's real intention. As the first step for doing this task, I briefly survey the theoretical background and the basic principles of scalar inference, especially focusing on Horn's and Matsumoto's scales (which is viewed as a complement to Horn's scale).

2. What is Scalar Inference (including Scalar Entailment and Scalar Implicature)?

2.1. The Gist of Horn' s Scale

Horn's theory of scalar inference is based on Grice's conversational maxims. For example, Grice (1975)'s sub-maxim Q_1 from the Maxim of Quantity is 'Make

your contribution as informative as is required' and the hearer-oriented, while the sub-maxim Q_2 is 'Do not make your contribution more informative than is required' and the speaker-oriented. Similarly, considering the Maxim of Manner, 'Avoid obscurity and ambiguity' concerns the hearer's position, whereas 'Be brief' does the speaker's.

From this viewpoint, Horn (1989, 2004) reconstructs all the Gricean maxims as two major principles, except the Maxim of Quality which is the most fundamental, and shows an inferential model in accordance with these: the Q principle (the hearer-oriented) and the R principle (the speaker-oriented). First of all, consider his Q principle and R principle (Horn 2004):

a. Q Principle (Hearer based): MAKE YOUR CONTRIBUTION SUFFICIENT.
Say as much as you can (modulo Quality and R)
Lower-bounding principle, including upper-bounding implicata
(It collects Grice's first Quantity maxim along with the first two 'clarity' sub-maxims of Manner.)

b. R Principle (Speaker based): MAKE YOUR CONTRIBUTION NECESSARY.
Say no more than you must (modulo Q)
Upper-bounding principle, including lower-bounding implicata
(It collects Grice's second Quantity maxim, Relation maxim and the last two sub-maxims of Manner.) (Horn 2004: 13)

According to Horn (2004), Q principle is related to scalar inference among the expressions where quantitative grades are set, while R principle is connected to an extended interpretation for the content of the utterance. Consider one example connected to Horn's R Principle:

(5) We went to the zoo yesterday. The elephant was sick.
↦ 'The elephant in the zoo where we went yesterday was sick'.

In (5), the hearer may infer that 'the elephant that the speaker has mentioned belongs to the zoo to which they went, and it was sick', even though the speaker does not add any extra information to his original utterance. In this

case, R Principle claiming 'Say no more than you must' works, because the speaker and the hearer assume that they should refer to the mutually most relevant elephant.

After Horn, Levinson (2000) developed a neo-Gricean account that re-thought the maxims as the Q[uantity]-Principle, I[nformativeness]-Principle and M[anner]-Principle. Levinson seeks to make a clearer distinction between semantic minimization ('semantically general expressions are preferred to semantically specific ones') and expression minimization ('"shorter" expressions are preferred to "longer" ones') than in Horn's approach.

(6) a. **The Q-Principle** (simplified)
Speaker: Do not say less than is required (bearing the I-Principle in mind).
Addressee: What is not said is not the case.

b. **The I-Principle**
Speaker: Do not say more than is required (bearing the Q-Principle in mind).
Addressee: What is generally said is stereotypically and specially exemplified.

c. **The M-Principle**
Speaker: Do not use a marked expression without reason.
Addressee: What is said in marked way is not unmarked.
(Levinson 2000: 75-164)

Thus, Levinson distinguishes between pragmatic principles governing an utterance's surface form and pragmatic principles governing its information content (Huang 2007: 41).

By Q principle, the speaker makes an utterance as informative as possible; if his utterance contains a weaker content, that is, not the strongest one, the hearer might think that the speaker implicates 'it is not the strongest content.' Therefore, Q principle can be viewed as the evidence confirming that language users' observe the Maxim of Quantity in the process of communication, and scalar inference (including scalar entailment and scalar implicature) is based on this Q principle. According to Q principle, when a

'scale' is formed, from the strongest one to the weakest one, the graded expressions are arranged in order of their semantic strength as a basic criterion of forming the scale. Scalar implicature is one of the most representative ways of pragmatic inference based on Horn's scale. In relation to this point, consider the following typical examples of scalar implicature:

(7) Scale: <always, usually, often, sometimes>
Joe sometimes plays Stacraft game alone.
↦ 'Joe does not {always / usually / often} play Stacraft game alone'.

(8) Scale: <must, can>
They can leave tomorrow.
↦ 'It is not the case that they must leave tomorrow'.

Levinson (1983) formularizes a rule eliciting scalar implicature as follows:

(9) **Scalar implicatures**:
Given any scale of the form $<e_1, e_2, e_3, \cdots e_n>$, if a speaker asserts $A(e_2)$, then he implicates $\neg A(e_1)$[2], if he asserts $A(e_3)$, then he implicates $\neg A(e_2)$ and $\neg A(e_1)$, and in general, if he asserts $A(e_n)$, then he implicates $\neg(A(e_{n-1}))$, $\neg(A(e_{n-2}))$ and so on, up to $\neg(A(e_1))$. (Levinson 1983: 133)

However, the notion of 'scale' is necessary for explaining not only implicature but also entailment as discussed in chapter 1; namely, the expressions forming a scale are put in the relation of logical entailment as well. Thus, I modify Levinson's generalization as follows:

(10) **Scalar entailment and scalar implicature**:
Given any scale of the form $<e_1, e_2, e_3 \cdots e_n>$, i) if a speaker asserts a sentence including (e_{n-1}), it entails a sentence including (e_n); ii) uttering a sentence including (e_n) implicates the negation

2) The symbol ¬ means 'negation'.

> of a sentence including (e_{n-1}). In other words, a sentence $S(e_{n-1})$ serially entails $S(e_n)$, and the sentence $S(e_n)$ serially implicates $\neg(S(e_{n-1}))$.

For instance, talking about fuel efficiency of a car, we can assume this scale <first grade, second grade, third grade, fourth grade, fifth grade>. When somebody says, "The fuel efficiency of this car is the third grade," it serially scalar-implicates from 'It's not the second grade' to the negation of the strongest expression in the whole scale. Similarly, the strongest expression, *the first grade*, serially entails from weaker one to the weakest one in the same scale.

2.2. Some Explanatory Lacunae in Horn's Scale by Q-Principle

In spite of the clear generalization of Horn's scale by Q principle, however, there are some explanatory problems of this theory, which cannot naturally be applied in ordinary language use. The first problem comes from the matter of interpreting scalar entailment. For instance, according to the logic of scalar entailment, if 'Tom has two children' is true, then 'Tom has one child' should also be true. But, if I serve a meal for only three persons, when I invite Tom's family (composed of two adults and two children) to my home, does it make sense? In this case, the host is to remember that 'two children' is literally 'two children.'

Similarly, Lee (2001: 235-236) points out another explanatory problem in Horn's scale. According to Lee, assuming a scale <boiling, hot, warm, chilly, cold, freezing>, it is impossible to explain why the utterance, "It was hot yesterday" should scalar-entail only 'It was warm yesterday,' not the weaker scales such as 'chilly,' 'cold' and 'freezing'. Moreover, we cannot say that "It was freezing yesterday" scalar-implicates 'It was not hot yesterday'. If it is the case that "It was freezing yesterday" scalar-implicates 'It was not hot yesterday', the conversational implicature from the original utterance should be cancelable. However, as 'It was freezing yesterday; in fact, it was hot yesterday' cannot be accepted, the implicated meaning from the original utterance is not cancelable; thus, those expressions in the scale of <boiling, hot, warm, chilly, cold, freezing> are not in the relation of conversational

implicature. For solving this explanatory problem, Lee (2001: 235) proposes a pragmatic constraint on forming scales:

(11) **Constraint on forming scale**: The expressions in a scale should share the same quality of positiveness or negativeness.
(Lee 2001: 235)

According to Lee's constraint in (11), *boiling, hot* and *warm* cannot belong to the same scale with *chilly, cold* and *freezing*. In other words, as *boiling, hot* and *warm* do not share the same quality with *chilly, cold* and *freezing*, they should be in separated scales; if so, it is possible to explain the problem discussed above.

Furthermore, I point out other explanatory problems due to merely applying Horn's scale, with reference to Horn's notion of 'metalinguistic negation'. Consider the following discourse between Oliver and Jenny, who are in love with each other:

(12) Jenny: (With a serious expression) "I don't like you, Oliver".
Oliver: (Very Surprised) "What?" "What on earth do you mean by that?" "Don't you like me?"
Jenny: "No, I don't like you". (With a big smile) "I love you!"

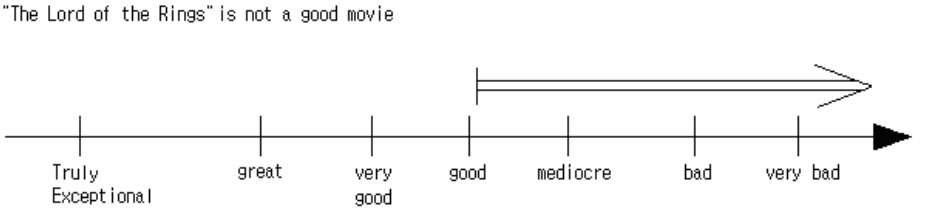
When a scale of <love, like> from Oliver and Jenny's discourse is formed, following the principle of scalar inference, *love* scalar-entails 'like', and *like* scalar-implicates '¬love'. Here, Jenny's last utterance is connected to the notion of 'metalinguistic negation'. In other words, what she said (i.e. "I don't like you" / '¬like') naturally scalar-implicates 'love' in accordance with the principle of scalar implicature that 'e_n implicates ¬(e_{n-1})'. Moreover, she makes her boyfriend happy by uttering this follow-up sentence, "I love you," lest him should misunderstand her real intention.

In relation to metalinguistic negation, it should be noted that this type of negation is used to negate the implicated meaning from the original affirmative sentence, and set a stronger point on the given scale, without negating the propositional content of the sentence or changing the original truth value of it. On the other hand, however, logical negation changes the truth value of the

original sentence. In other words, when logical negation is used in an affirmative sentence, it refers to a scalar point (or expression) whose semantic strength is weaker or an opposite side, by overturning the original truth value of the sentence.

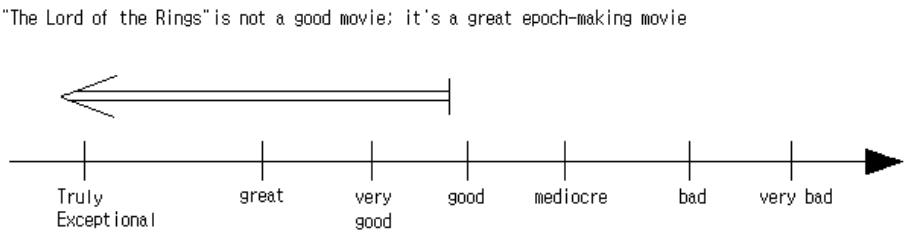
Judging from the point discussed above, metalinguistic negation can be good evidence supporting the pragmatic validity of scalar implicature. The figures given below show the difference between logical negation and metalinguistic negation; the former is logical negation, and the latter is metalinguistic one.

(13) **logical negation**:



(Lee 2002: 99)

(14) **metalinguistic negation**:



(Lee 2002: 100)

Returning to the discourse between Oliver and Jenny, from Jenny's utterance, it is possible to form a scale <love, like>, and according to the principle of scalar inference, *love* scalar-entails 'like', and *like* scalar-implicates 'not love'. Thus, if 'I love you' is true, then 'I like you' should also be true by scalar-entailment (i.e. When X is true, Y is also true), and 'I like you' should scalar-implicate 'I don't love you'. If so, when someone says, "I don't like you", this utterance should

scalar-entail "I don't love you" (i.e. When Y is false, X is also false) and scalar-implicate 'I love you' under the assumption that the speaker uses metalinguistic negation.

However, sometimes the conclusion of strictly applying the linguistic principle of scalar inference can be away from ordinary language users' normal intuition particularly in their colloquial language use. Considering a scale of <love, like>, the semantic difference between these two words can be neutralized as follows:

(15) Tom: What about going for a walk together?
Jerry: a. I'd like to.
b. I'd love to.

In (15), when Tom suggests that Jerry should go for a walk together, Jerry can accept Tom's suggestion by saying (15a) or (15b). However, even if Jerry says "I'd like to", instead of "I'd love to", Tom might not obstinately or stubbornly interpret Jerry's answer as the scalar-implicated meaning of 'Jerry does not love (want) to go for a walk together'.

Similarly, assuming a scale <murder, kill>, the strict semantic and pragmatic distinction between these two words can also be neutralized in ordinary language use. In relation to this point, consider the following example:

(16) Many people were killed by the enemy in the war.

Following the generalization by the principle of scalar inference, it is possible to expect that *murder* scalar-entails 'kill', and *kill* scalar-implicates 'not murder'. Thus, (16) should scalar-implicate 'Many people were not murdered by the enemy in the war'. However, considering that the nature of war always involves 'legal massacre', in which intentionality is always premised, *kill* in (16) is naturally interpreted as 'murder' although there is not any extra follow-up utterance, "but not murdered". Additionally, another example (15) can also support this point of view:

(17) Mary burst into tears when she received the list of the K.I.A. (i.e. 'Killed in Action') in the Vietnam War.

Many times, *the Killed in Action* in (17) means 'the Murdered in Action'; and English native speakers usually use *the K.I.A*, instead of *the M.I.A.* (i.e. 'Murdered in Action'). In fact, the intial of *M.I.A.* generally means 'Missing in Action', rather than 'Murdered in Action'.

The matter of politeness can also be another factor that eclipses the conclusion of strictly applying the principle of scalar implicature by Horn. According to Leech (1983), if different maxims collide with each other, the maxim of politeness is prior to others. The relevant maxim held by Leech (1983) is as follows:

(18) **Principle of Politeness: The Approbation Maxim**
"Avoid saying unpleasant things about others, and more particularly, about the hearer". (Leech 1983: 135)

With reference to Leech's principle given in (18), suppose that a baseball coach says, "Your record is a little bit deficient today" to his young player whose record is much worse than that of ordinary times. According to the principle of scalar implicature, forming a scale <very deficient, a little bit deficient>, once the coach utters "Your record is a little bit deficient today", his utterance should scalar-implicate 'your record is not very deficient'. However, (19) is not acceptable:

(19) "Your record is a little bit deficient".
↦ '(The speaker believes) that the hearer's record is not very deficient'.

In (19) the utterance intends to save the hearer's face, and from Brown and Levinson's (1987) point of view, this is an example of not threatening the hearer's positive face; so the approbation maxim (of principle of politeness) is prior to Grice's Q_1 maxim, without reference to the principle of scalar implicature.

2.3. A Complement to Horn's Scale: Matsumoto (1995)'s Scale

In the previous sections 2.1 and 2.2, I discussed how the utterances connected to the notion of 'scale' which describes slightly different situations can be interpreted. Here, the criterion of forming a scale is the difference of the semantic strength which each member in the same scale has. For instance, there is a difference of the semantic strength between "It's hot" and "It's warm"; I call this *scale of situation*. In a scale of situation, scalar implicature by Grice's Q_1 maxim normally occurs like 'It's warm today ↦ 'It's not hot today', or 'It's not boiling today'.

However, assuming a scale of <Baker Street 221B, London, England>, we can see that Baker Street 221B is a part of London, and London is a part of England. Here, this scale consists of meronymy (which describes the semantic relation of 'the part vs. the whole'). If Tom asks Jerry where Spike studies pragmatics, then Jerry might say, "Spike studies pragmatics in London". But Jerry's utterance cannot scalar-implicate 'not in Baker Street 221B'. Therefore, this type of scale is different from Horn's scale exploiting Q principle, discussed in the previous sections. Unlike Horn's scale by Q principle, formed by the difference of semantic strength describing the given situations (which I call *scale of situation*), forming this new type of scale (which I call *scale of specificity*) depends upon how specifically the given situation is described; in this case, Horn's R principle for extended interpretation is to be applied, instead of Q principle. In other words, whatever the expression the speaker chooses in the given scale of specificity, he may describe the given situation and the truth value of his utterance is true. However, although the speaker chooses a weaker expression W, it does not scalar-implicate the negation of a stronger expression S, unlike the case of 'scale of situation'. In relation to this matter, Matsumoto (1995) proposes the notion of 'Conversational Condition', and holds that a scale does not license a Quantity-1 implicature if this condition is violated:

(20) **Conversational Condition**: The choice of W instead of S must not be attributed to the observance of any information-selecting Maxim of Conversation other than the Quality Maxims and the Quantity-1 Maxim (i.e. the Maxim of Quantity-2, Relation, and Obscurity,

Avoidance, etc.) (Matsumoto 1995: 25)

In the next section, I analyse a different type of scalar implicature on the basis of three sub-conditions of the Maxim of Quantity-2, Relation, and Obscurity Avoidance.

2.3.1. The Quantity-2 Condition

Matsumoto (1995) defines the Quantity-2 Condition as follows:

(21) **The Quantity-2 Condition**: S must not convey more information than is required in the particular context of utterance in which W is used. (Matsumoto 1995: 27)

The reason why Jerry chooses *London* instead of *Baker Street 221B* is not because he observes the Q-2 condition but because he intends to utter what the given context level requires, and this point is connected to Grice's second Maxim of Quantity (i.e. 'Do not make your contribution more informative than is required') and Horn's R principle based on it. Matsumoto (1995) holds that there are two kinds of quantity information: 1) quantity on the horizontal axis and 2) quantity of the vertical axis. The former (i.e. scale of situation) is the semantic strength of information on physically or socially defined scale such as quantity, temperature, age, height, military rank and so on, whereas the latter (i.e. scale of specificity) refers to the degree of the detailedness or specificity of information, as which a referent or a state is described. The Quantity-2 Condition includes the information on the vertical axis. Thus, in this case, R principle is applied, instead of Q principle. I take an example in order to explain the notion of 'scale of specificity'. Consider the following example:

(22) a. (Both Tom and Jerry know that there are only royal galas[3] in Spike's orchard.)
Tom: What is Spike doing in his orchard?
Jerry: He is picking apples there.

3) Royal gala is a variety of apple.

b. ¬(↦ ('Tom believes that) what Spike is picking are not royal galas'.)

Matsumoto (1995) explains this phenomenon by citing the position of Hirschberg (1985: 160) as follows:

> Unless certain conditions obtain, "salient scales of specificity in taxonomy (which function as Horn scale in context) are upper-bounded by the basic level term, with more specific terms excluded from a salient scale. In the present account, this failure to produce implicatures can be attributed to the violation of the Quantity-2 Condition. (Matsumoto 1995: 29)

Thus, the reason why the implicature like (22b) does not occur is because Quantity-2 Condition is not observed. In other words, as this context does not require any information about a variety of apple, Jerry's utterance expressing the state of a basic level is fully informative for fulfilling the intended inference. However, consider another example similar to but slightly different from the example in (22):

(23) a. (Tom whose mother hates all reptiles secretly bought an iguana in a pet shop; and Jerry knows it. One day, Tom's mother who knows that Jerry went to the pet shop with Tom asks Jerry what her son bought there.)

Tom's mother: What did Tom buy in the pet shop?
Jerry: Well, I saw him buy just a small animal there.

b. ¬(↦ (Tom's mother believes that) Jerry does not exactly know what Tom bought.)

Assuming a scale <iguana, animal>, unlike the example in (22), even though this context in (23) requires the exact information about the species of the animal which Tom bought in the pet shop, Jerry intentionally uses the weaker expression W, instead of the stronger expression S. Thus, Jerry's scalar utterance expresses that the state of a basic level is not fully informative for fulfilling the

intended inference. Why does Jerry intentionally say so? Jerry's utterance is involved in his non-linguistic motivation, rather than the pure linguistic factors. In other word, if Jerry tells Tom's mother the truth, Tom might misunderstand that Jerry snitched on to Tom's mother the fact that Tom bought an iguana without his mother's permission, even though his mother hates all reptiles. This could bring about potential conflicts between Tom and Jerry in the future, and Jerry wants to avoid any potential conflicts with Tom, cooperating in the given conversational situation. This point clearly shows that ordinary language users' inferential process can be influenced by other various non-linguistic factors such as 'politeness' and 'face-saving strategy', besides linguistic rules and principles.

2.3.2. The Relevance Condition

The second example of Matsumoto's conversational condition is connected to the Maxim of Relation:

(24) **The Relevance Condition**: the information that S conveys must be relevant to the discourse in which W is used. (Matsumoto 1995: 37)

Considering a scale <British Prime Minster, the Lord Mayor of London>, Tom's utterance in the following conversation implicates that 'Jerry's uncle was not British Prime Minster':

(25) a. Tom: "What did your uncle do?"
Jerry: "He was a politician. He was the Lord Mayor of London".
b. ↦ '(Jerry believes that) his uncle was not British Prime Minister'.

On the other hand, in (26) even though Jerry's answer is the same, if the content of Tom's question is changed, Jerry's answer cannot implicate that 'Jerry's uncle was not British Prime Minister':

(26) a. Tom: "What city was your uncle mayor of?
Jerry: "He was the Lord Mayor of London".
b. ¬(↦ '(Jerry believes that) his uncle was not British Prime Minister'.

Adapting the point of view in Matsumoto (1995: 39), this phenomenon can be explained as follows: the implicature in (25b) comes from a scale composed of rank terms (i.e. *British Prime Minster, the Lord Mayor of London*). The pragmatic difference between (25) and (26) presents that this scale brings about an implicature only when the point of the utterance is to inform the hearer of the highest office the person in question held. In other words, the Relevance Condition is satisfied in (25), and the information conveyed by *British Prime Minister* is relevant to the present discourse; whereas in (26) the same condition is not observed and the information carried by the same expression (i.e. *British Prime Minister*) is not relevant to it.

However, Lee (2001) points out that it is impossible to explain why the following utterance like (27) does not license any implicature by merely applying Matsumoto's Relevance Condition:

(27) Who do you think will vote for Tony Blair?
  a. I think the Lord Mayor of London will vote for Tony Blair.
  b. I think British Prime Minister will vote for Tony Blair.

According to Lee (2001: 244), both (27a) and (27b) can be appropriate answers for the question (27), because British Prime Minister can be a subject of the subordinate clause, instead of the Lord Mayor of London; thus, (27a) does not violate the Relevance Condition. However, it is not the case that (27a) implicates the negation of (27b). Lee (2001) holds that there does not exist any scale itself here. Moreover, according to Matsumoto (1995), <British Prime Minster, the Lord Mayor of London> is viewed as a scale; if so, (28) with S should entail (29) with W. However, the result is far from Matsumoto's viewpoint, as follows:

(28) I saw British Prime Minister.
(29) I saw the Lord Mayor of London.

For solving this explanatory problem, Lee (2001) proposes 'Constraint by the Same Criterion' (which he calls). According to him, a scale like <British Prime Minster, the Lord Mayor of London> is formed not so much by

semantic strength as by a pragmatic criterion such as 'a politician with the highest rank, who is expected to vote for Tony Blair'. That is why it is not possible to explain why (28) cannot entail (29) until the pragmatic criterion is provided, instead of applying semantic strength. In section 3.2, I discuss Lee's 'Constraint by the Same Criterion' in detail, which is an alternative for explaining the lacuna of Matsumoto's 'Relevance Condition'.

There is one more case in which Quantity-1 implicature is not evoked. If the speaker chooses W instead of S, sometimes it might be due to observing the Maxims of Manner (i.e. Obscurity Avoidance or Brevity), rather than the principle of Quantity-1. In the next section, I discuss this point.

2.3.3 The Non-Obscurity Condition

Matsumoto (1995) holds that besides two sub-instances of the Conversational Condition based on the Maxim of Quantity-2 and Relation, there is two more sub-instances connected to the Maxims of Manner: 1) the Maxim of Obscurity Avoidance (i.e. 'Avoid obscure expressions') and 2) the Maxim of Brevity.[4] These two sub-instances compose the Non-Obscurity Condition, and it is defined as follows:

(30) **The Non-Obscurity Condition**: S must not be obscure (to the hearer).

In relation to the Maxim of Obscurity Avoidance which is one of the two components of the Non-Obscurity Condition, Schegloff (1971) holds that when the speaker describes a location, he does not usually mention a proper name (without an explanation or qualification of some sort) if the hearer is not exactly aware of the name, and a Quantity-1 implicature does not occur here. Consider the following example:

(31) a. Spike: "What town will Tom visit?"
Jerry: "He will visit a small town not far from London".

4) In relation to other Maxims of Manner (i.e. 'Be orderly' and 'Avoid Ambiguity'), according to Matsumoto's point of view, it is not necessary to discuss them, 'because it is not clear how they relate to the choice of S vs. W' (Matsumoto 1995: 39).

b. 'Jerry does not know which of the small towns not far from London Tom will visit'.

In (31a), Jerry is to tell Spike the exact name of a small town which Tom will visit; but Jerry's answer is not fully informative. According to Schegloff (1971)'s viewpoint, if Jerry thinks that Spike well knows the names of particular towns near London area, then (31b) is implicated. On the other hand, however, if Jerry thinks that Spike is not familiar with the nearby area of London, (31b) is not implicated.

Another sub-instance involved in the Non-Obscurity Condition is the Maxim of Brevity, which is based on 'Be brief'. Matsumoto (1995: 42-43) holds that "This maxim states that when there are two roughly synonymous expressions one of which is apparently more prolix than the other, the speaker who uses the prolix expression implicates that the intended meaning is distinct from that conveyed by the briefer expression". Levinson (1983, 2000) proposes 'Brevity Condition' as follows:

(32) **Brevity Condition**: The stronger item must be of equal brevity to the weaker item. (Levinson 1983: 135)

The relevant example is as follows:

(33) a. "Watch out for that spider".
b. ¬(↦ 'The speaker does not know the color, size, or exact position of the spider he warns about'.
c. "Watch out for the black, half-inch long spider that has a green dot in its center and is about six inches from your left shoulder at a vertical angle of about sixty degrees". (Matsumoto 1995: 43)

In (33), saying (33a), which is less informative than (33c), does not implicate (33b). In other words, the speaker utters (33a) not because he does not know the exact information about the spider he warns about but because further details about the spider are not necessary in the given situation and he just needs to urgently tell the hearer the danger in a briefer expression. Thus, in the

discourses like (33), Q-2 principle (related to Horn's R principle) works and licenses implicature, instead of Q-1 principle.

3. Forming Scales by a Pragmatic Criterion and a Constraint on it

3.1 Forming Scale by a Pragmatic Criterion

In the previous chapter, I surveyed the gist of Horn's scale based on the notions of 'semantic strength (of situation)' and 'Quantity-1 principle', and discussed Matsumoto's scale which focuses on 'specificity' and 'Quantity-2 principle', which can be viewed as a complement to Horn's. Furthermore, I pointed out that sometimes interpreting scalar inferential utterances involves non-linguistic factors and strictly applying the principle of scalar inference might bring about failure of reaching the real conclusion intended by the speaker.

In this chapter, I discuss forming a scale by a pragmatic criterion and a constraint on it; and this matter is significant in that it can be the evidence showing that sometimes non-linguistic factors may influence interpreting the meaning of scalar implicature. According to Horn (1989), a scale is often formed by a pragmatic criterion. For example, Bill Gates, Richard Branson and David Beckham share the point that they are all famous and rich. Thus, it is possible to form a scale like <Bill Gates, Richard Branson, David Beckham>. In relation to this scale, consider the following example:

(34) Although Tom is not as rich as Bill Gates, he is as rich as Richard Branson or at least David Beckham.

However, if this order is changed, it sounds odd; so it means that there is a scalar hierarchy, as a pragmatic criterion of 'famous and rich men', between them. Consider another example in (35).

(35) ?? Although Tom is not as rich as Richard Branson or David Beckham, at least he is as rich as Bill Gates.

If so, it is possible to form a scale of <Bill Gates, Richard Branson, David Beckham> by the pragmatic criterion of 'famous and rich men in the contemporary age'. Now it is necessary to confirm whether or not this scale licenses scalar inference (including scalar entailment and scalar implicature) by applying the principle in (10), which was discussed in section 2.1. I recall the principle given in (10):

(36=10) **Scalar entailment and scalar implicature**:

> Given any scale of the form $<e_1, e_2, e_3 \cdots e_n>$, i) if a speaker asserts a sentence including (e_{n-1}), it entails a sentence including (e_n); ii) uttering a sentence including (e_n) implicates the negation of a sentence including (e_{n-1}). In other words, a sentence $S(e_{n-1})$ serially entails $S(e_n)$, and the sentence $S(e_n)$ serially implicates $\neg(S(e_{n-1}))$.

First of all, the utterance (37) including the strongest expression *Bill Gates* in this scale can serially entail both (38) and (39):

(37) Bill Gate is not rich enough to buy ham sandwiches for all the people in China and India.
(38) Richard Branson is not rich enough to buy ham sandwiches for all the people in China and India.
(39) David Beckham is not rich enough to buy ham sandwiches for all the people in China and India.

However, it is not possible to say that if (40) is true, (41) is also true; so scalar-entailment does not always occur in this scale:

(40) Richard Branson sold his luxurious sports car.
(41) David Beckham sold his luxurious sports car.

Furthermore, we cannot see the utterance including a weaker expression W scalar-implicates the negation of the utterance including a stronger expression S; namely, nobody infers (43) or (44) from (42).

(42) David Beckham frequently quarrels with his wife nowadays.
(43) Richard Branson does not frequently quarrel with his wife nowadays.
(44) Bill Gates does not frequently quarrel with his wife nowadays.

Here, the criterion forming the scale is neither Horn's 'semantic strength' nor Matsumoto's 'specificity'; the criterion comes from a pragmatic factor (i.e. ordinary language users' knowledge or common sense about the world). Thus, Lee (2001) points out that not always a scale involving a pragmatic criterion licenses scalar-inference connected to Q principle by Horn, and it is necessary to pursue a constraint on forming scales by this principle.

3.2. Constraint by the Same Criterion

Judging from all the examples examined in section 3.1, even though the principle of (36=10) exactly predicts the inferential relations in various scales which have been discussed, it cannot successfully be applied to a scale such as <Bill Gates, Richard Branson, David Beckham> pragmatically formed. For this reason, it is necessary to pursue an appropriate constraint on this type of scale. In other words, this is just a scale of 'famous and rich men in the contemporary age', not that of other criteria such as 'the men who recently sold their luxurious sports cars' or 'the men who frequently quarrel with their wives nowadays.' Thus, a scale like <Bill Gates, Richard Branson, David Beckham> cannot be formed by other criteria.

However, the principle of (36=10), once a scale is semantically or pragmatically formed, predicts that when other parts of the sentence are fixed, scalar entailment or implicature always occur without reference to the criterion of forming scale. In other words, while a scale semantically determined always sets the inferential relation of the sentence in accordance with a fixed principle of (36=10), a scale pragmatically formed does not. In the latter, unlike a scale semantically set, the semantic cohesion between the expressions forming the scale is confined to only one criterion, and only when the sentence connected to that criterion is used, scalar inference (including scalar-entailment and scalar-implicature) may occur. Thus, Lee (2001) holds that in the process of inferring a scale pragmatically determined, the following constraint is required:

(45) **Constraint by the Same Criterion**: In the inferential process of calculating scalar entailment or scalar implicature given in (36=10), other criteria, which are not the original criteria made at the outset, cannot be applied.[5] (Lee 2001: 242)

Lee (2001: 244) claims that this constraint can explain why (46a=27a) cannot scalar-implicate the negation of (46b=27b), which was discussed in section 2.3.2; furthermore, why (47=28) cannot scalar-entail (48=29). Consider the following examples:

(46=27) Who do you think will vote for Tony Blair?
a. I think the Lord Mayor of London will vote for Tony Blair.
b. I think British Prime Minister will vote for Tony Blair.
(47=28) I saw British Prime Minister.
(48=29) I saw the Lord Mayor of London.

Following Lee's point of view, when a pragmatic scale is set, there should be a relevant criterion which enables the scale to be formed. Assuming a scale of <British Prime Minister, the Lord Mayor of London>, this scale can be formed by a criterion such as 'the politician in the highest rank that is expected to vote for Tony Blair'; if so, (46a=27a) can scalar-implicate the negation of (46b=27b). However, Lee points out that the question in (46=27) is not enough to make the hearer form this scale by such criterion, because the question in (46=27) offers only the criterion of 'the people who are expected to vote for Tony Blair'.

Likewise, this constraint is also valid for explaining why (47=28) cannot scalar-entail (48=29). In other words, as (47=28) and (48=29) are used under another criterion of 'the politician whom I saw', not 'the people who are expected to vote for Tony Blair', any scalar entailment is not licensed.

However, it is still questionable that without fully considering non-linguistic factors such as language users' common sense and intuition, merely applying the various principles of scalar inference discussed in the previous sections can really cover every pragmatic facet of our ordinary language use. If so, the question is how much language users should take non-linguistic factors into

5) This is my translation from Korean to English.

account in the process of communication connected to scalar inference. In fact, scalar inference is one of the most pragmatic phenomena, and pragmatics which considers non-linguistic factors is optimized to deal with this matter.

Nevertheless, the point discussed above is nothing less than the problem making language users undergo difficulties about reaching the conclusion. If a scale is formed by a pragmatic criterion, it means that interpreting the sentence by a pragmatic scale may go beyond the linguistic dimension. For instance, recalling a scale of <Bill Gates, Richard Branson> formed by the pragmatic criterion of 'famous and rich men in the contemporary age' in section 3.1, if someone says, "Richard Branson is not rich enough to buy ham sandwiches for all the people in China and India", it should scalar-implicate 'Bill Gate is rich enough to buy ham sandwiches for all the people in China and India' (i.e. the negation for the stronger expression in the same scale).

However, the interpretation proposed above might be controversial as long as ordinary language users consider their common sense in the world; because most people believe that 'no matter how rich Bill Gates is, it is impossible for him to buy ham sandwiches for all the people in China and India'. But in fact, this is nothing but ordinary language users' belief based on their common sense; strictly speaking, nobody knows whether or not Bill Gates can really afford to buy ham sandwiches for all the people in China and India. If the speaker says "Richard Branson is not rich enough to buy pencils for all the people in China and India", it might be more probable that this utterance easily licenses the implicature of 'Bill Gate is rich enough to buy pencils for all the people in China and India, although Richard Branson is not'. But, if the speaker says "Richard Branson is not rich enough to buy Ferrari sports cars for all the people in China and India", nobody would think that this utterance scalar-implicate 'Bill Gate is rich enough to buy Ferrari sports cars for all the people in China and India'. Thus, this conclusion depends upon ordinary language users' non-linguistic knowledge, rather than the pure linguistic factors. Of course, some people might argue that this expression can be acceptable as long as it is regarded as hyperbole; however, if so, we should admit that this interpretation gets out of the pure domain of scalar inference.

4. Conclusion

In this paper, I discussed the nature of scalar inference and pointed out some explanatory lacunae of Horn's scale formed by Q-1 principle (i.e. Q principle which is hearer based). Focusing on 'semantic strength' and 'situation', sometimes Horn's approach fails to exactly interpret other types of scalar expressions involved in non-linguistic (particularly socio-cultural) factors. By examining some examples, I claimed that when the hearer interprets the speaker's scalar utterance, it is necessary to fully consider other non-linguistic factors in order to grasp the speaker's real intention, as confirmed by the positions of Leech (1983) and Brown and Levinson (1987).

Furthermore, I surveyed Matsumoto's scale which focuses on 'specificity' instead of 'semantic strength'. Unlike Horn's scale which exploits Q-1 principle and focuses the semantic strength of the expressions in a scale, Matsumoto's scale uses Q-2 principle (i.e. R principle which is speaker based) and considers 'specificity'. Thus, Matsumoto's scale is a meaningful complement to Horn's scale in that it explains how specifically the situation is described, which Horn's scale explaining how much the semantic strength of the expressions in a scale influences the scalar inference misses out.

For some explanatory problems of Matsumoto's scale, Lee (2001) proposes a complementary approach (i.e. 'Constraint by the Same Criterion') to them. By applying this approach, it is possible to solve the explanatory problems which even Matsumoto's approach cannot explain.

Finally, under the support of Lee (2001)'s approach, I confirmed that it is necessary to pursue a constraint on forming scale. Lee's 'constraint by the same criterion' holds that unless interpreting a scale formed by a pragmatic criterion is restricted by the original criterion applied at the outset, it is impossible to treat some cases of scalar inference involving pragmatic criteria. However, as this constraint is based on ordinary language users' intuition and common sense, rather than a pure linguistic theory, no matter how excellent the linguistic theory is, without fully considering these non-linguistic factors, it is not possible to reach the real conclusion intended by the speaker.

References

Brown, P. & Levinson, S. (1987). *Politeness: Some universals in language use*. Cambridge: Cambridge University Press.

Gazdar, G. (1979). *Pragmatics: Implicature, Presupposition, and Logical From*. London: Academic Press.

Grice, H. P. (1975). Logic and Conversation. In Martinich, Aloysius P. (ed.) *The Philosophy of Language* (third edition). Oxford: Oxford University Press. 156-167.

Hirschberg, J. (1985). *A Theory of Scalar Implicature*. Ph.D. dissertation. University of Pennsylvania.

Horn, L. (1985). Metalinguistic Negation and Pragmatic Ambiguity. In *Language* 61: 121-174.

Horn, L. (1989). *A Natural History of Negation*. Chicago: The University of Chicago Press.

Horn, L. (2004). Implicature. In Horn, Laurence R. and Ward, Gregory (eds.), *The Handbook of Pragmatics*. Oxford: Blackwell. 3-28.

Huang, Y. (2007). *Pragmatics*. Oxford: Oxford University Press.

Lee, Sungbom. (2001). *Churoneui Whayongron* [Pragmatics of Inference]. Seoul: Hankuk Munhwa-sa.

Lee, Sungbom. (2002). *Yeong-eh Whayongron* [*English Pragmatics*]. Seoul: Hanhuk Munhwasa.

Leech, G. (1983). *Principles of Pragmatics*. New York: Longman.

Levinson, S. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, S. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge: MIT Press.

Matsumoto, Y. (1995). The Conversational Condition on Horn Scale. In *Linguistics & Philosophy 18*: 21-60.

Schegloff, E. (1971). Notes on a Conversational Practice: Formulating Place. In D. Sudnow (ed.). *Studies in Social Interaction*. New York: Free Press. 71-119.

Dae-Young Kim
Department of English Education
Jeonju University
303 Cheonjam-ro, Wansan-gu
Jeonju 560-759, Korea
Email: amante516@jj.ac.kr

Received on October 27, 2016
Revised version received on December 19, 2016
Accepted on December 30, 2016