# Partial Blocking, Associative Learning, and the Principle of Weak Optimality

**Anton Benz**
Zentrum für Allgemeine Sprachwissenschaft, Berlin, Germany

## 1. Introduction

One of the selling–points of Bi–OT is its success in explaining partial blocking phenomena. In **(i)** it has to explain why *kill* tends to denote a *direct* killing whereas *caused to die* tends to denote an *indirect* killing [6]:

**(i)** **a)** Black Bart killed the sheriff.
**b)** Black Bart caused the sheriff to die.

The Bi–OT explanation is based on the principle of weak optimality, a generalisation of a rule known as *Horn's division of pragmatic labour* [10, p. 22]: Marked forms typically get a marked interpretation, and unmarked forms an unmarked interpretation. *Kill* is the less marked form, and if we assume that speakers prefer less marked forms over marked forms, then *kill* is the optimal way to denote a killing event. As direct killing is the normal and expected way of killing, the hearer should have a preference for interpreting the speaker's utterance as referring to a direct killing. We can see that *kill* and *direct killing* build an optimal form–meaning pair from both perspectives. In addition we can see that the marked form tends to denote the less expected meaning, i.e. *cause to die* tends to denote an *indirect killing*. In general, if $F_1$ and $F_2$ are forms and $M_1$ and $M_2$ are meanings where $F_1$ is preferred over $F_2$ and $M_1$ over $M_2$, then $F_1$ tends to denote $M_1$ and $F_2$ to tends denote $M_2$:

**(ii)**

$$\begin{array}{cccc} & M_1 & & M_2 \\ F_1 & \bullet & \longleftarrow & \bullet \\ & \uparrow & & \uparrow \\ F_2 & \bullet & \longleftarrow & \bullet \end{array}$$

Horn explains his principle by recursion to two pragmatic principles, called the Q– and R–principle. Blutner [5] gave them a formally precise formulation. Specifically, he made explicit the role of switching between speaker's and hearer's perspective. This laid the foundation for an optimality–theoretic reformulation, and thereby for placing radical pragmatics in the broader linguistic context provided by OT. In this paper we are going to explain partial blocking as the result of diachronic processes based on what we will call *associative learning*.

(1) Bi–OT over–generates partial blocking, i.e. it predicts partial blocking for many examples where blocking is not observable; (2) Bi–OT in its original form has only a weak foundation, i.e. there is no good explanation for the principle of weak optimality which does (a) not make an (implicit) appeal to Horn's principle of pragmatic labour, and (b) provides more than just an algorithm for how to calculate weakly optimal form–meaning pairs. Game theory has been proposed as a remedy for the last problem [9]. We will discuss Bi–OT at more length in Section 2, and in Section 3 we consider van Rooy's game–theoretic approach to explaining Horn's division of pragmatic labour [16]. Partial blocking can be observed in examples where expressions are unambiguous and where there would be an alternative form

for denoting the more marked meaning. We will see that these assumptions about language make van Rooy's model inapplicable.

Originally, Blutner understood his theory from a diachronic perspective‡. We take this idea more seriously. We claim that partial blocking can be explained as an effect of *associative learning* plus speaker's preferences on forms. It emerges as a result of a diachronic process. We explain Example **(i)** by postulating the following five stages: (1) In the initial stage all killing events are direct killing events. The speaker will always use *kill* to denote these events. (2) Interpreters will learn that *kill* is always connected with direct killing. They *associate kill* with direct killing. (3) The speaker will learn that hearers associate *kill* with direct killing. (4) If then an exceptional event occurs where the killing is an indirect killing, the speaker has to avoid misleading associations, and use a different form. In this case it is the more complex form *cause to die*. (5) The hearer will then learn that *cause to die* is always connected to an untypical killing. By *associative learning* we mean the learning process in (2), (3), and (5). We postulate the following principle related to the hearer:

> In every actual instance where the form $F$ is used for classifying events or objects it turns out that the classified event or object is at least of type $\mathbf{t}$, then the hearer learns to associate $F$ with $\mathbf{t}$, i.e. he learns to interpret $F$ as $\mathbf{t}$.

A similar principle is assumed for the speaker to explain step (3). Given a set of semantically synonymous expressions, how can associative learning and speaker's preferences lead to a change in interpretation? In Section 4 we work out a formal model which describes diachronic processes related to associative learning.

## 2. Bi–OT and Weak Optimality

According to OT, producer and interpreter of language use a number of constraints which govern their choice of forms and meanings. These constraints may get conflict with each other. OT proposes a mechanism for how these conflicts are resolved. It assumes that the constraints are ranked in a linear order. If they get into conflict, then the higher-ranked constraints win over the lower–ranked ones. This defines preferences on forms and meanings.

Optimality theory has divided into many sub–theories and variations. Beaver and Lee [2] provide for a useful overview of versions of optimality–theoretic semantics. They discuss seven different approaches. In particular they compare them according to whether they can explain partial blocking. It turns out that the only approach which can fully justify Horn's division of pragmatic labour is Blutner's Bi–OT [2, Sec. 7 and 5].

What are the structures underlying Bi–OT? In bidirectional OT it is common to assume that there is a set $\mathcal{F}$ of *forms* and a set $\mathcal{M}$ of *meanings* [6]. A set $\mathrm{Gen}$, the so–called *generator*, tells us which form–meaning pairs are grammatical. The grammar may leave the form–meaning relation highly underspecified. In a graphical representation like **(ii)** a grammatical form–meaning pair $\langle F, M\rangle$ is represented by a bullet at the point where the row for $F$ and the column for $M$ intersect. Underspecification means that a row corresponding to a form $F$ may contain several bullets. The speaker has to choose for his utterance a form which subsequently must be interpreted by the hearer. It is further assumed that the speaker has some ranking on his set of forms, and the hearer on the set of meanings. Blutner [6] introduced the idea that the speaker and interpreter coordinate on form–meaning pairs which are most preferred from both perspectives. The speaker has to choose for a given meaning $M_0$ a form $F_0$ which is optimal according to his ranking of forms. Then the interpreter has to choose for $F_0$ a meaning $M_1$ which is optimal according to his ranking of meanings. Then again the speaker looks for the

‡ Personal communication.

most preferred form $F_1$ for $M_1$. A form–meaning pair is optimal if ultimately speaker and hearer choose the same forms and meanings. If $\langle F, M\rangle$ is optimal in this technical sense, then the choice of $F$ is the optimal way to express $M$ so that both speaker's and interpreter's preferences are matched.

It is easy to see that the procedure for finding an optimal form–meaning pair stops for a pair $\langle F, M\rangle$ exactly if there are no pairs $\langle F', M\rangle$ and $\langle F, M'\rangle$ such that the speaker prefers $F'$ over $F$ given $M$ and the hearer prefers $M'$ over $M$ given $F$. In the graph **(ii)** $\langle F_1, M_1\rangle$ is optimal because there are no arrows leading from $\langle F_1, M_1\rangle$ to other form–meaning pairs. *Weak optimality* is a weakening of the notion of optimality. In **(ii)** we find that $F_2$ should go together with $M_2$. For $\langle F_1, M_2\rangle$ and $\langle F_2, M_1\rangle$ there is either a row or a column which contains it together with the optimal form–meaning pair $\langle F_1, M_1\rangle$. For $\langle F_2, M_2\rangle$ neither its row nor its column contains the optimal $\langle F_1, M_1\rangle$. If we remove the row and the column which contain $\langle F_1, M_1\rangle$, then $\langle F_2, M_2\rangle$ is optimal in the remaining graph. This can be generalised: If we remove from a given graph all rows and columns which contain an optimal form–meaning pair, then the optimal form–meaning pairs in the remaining graph are called *weakly optimal*. We can iterate this process until no more form–meaning pairs, and hence no graph, remains§.

*The Problem of Over–Generation* Bi–OT can successfully explain examples like **(i)** but if we apply it naively, then there are many examples where it over–predicts partial blocking. We first look at examples with anaphora resolution where it is semantically not clear who of the antecedents is male or female but where one of the alternatives is highly preferred. We don't get a marked interpretation for a marked expression‖:

**(iii)** **a)** The doctor kissed the nurse. She is really beautiful.
**b)** The doctor kissed the nurse. The woman is really beautiful.
**c)** The doctor kissed the nurse. Marion is really beautiful.
**d)** (?)The doctor kissed the nurse. SHE is really beautiful.

If the hearer has no special knowledge about the doctor and the nurse, he will interpret the second sentence as meaning *the nurse is really beautiful*. If we assume further that a pronoun is more economic than a proper name, and a proper name more economic than a definite description, then the speaker should continue his first sentence with *She is really beautiful*. The uses of *Marion* and *the woman* are less preferred, hence they should go together with a marked interpretation. If we apply the principle of weak optimality straightforwardly, then it predicts a tendency of e.g. *Marion*, or *the woman*, to indicate that the doctor is a woman. But for all three examples we get the same reference. If we stress the pronoun, then the sentence becomes ungrammatical rather than getting a marked reading.

Examples **(iii)** and **(iv)** are cases where underspecification is crucially involved. The next two examples represent cases without underspecification:

§ The principle of weak optimality is due to Blutner, see [5, 6]. He calls *superoptimality* what was later called weak optimality. The process for finding weakly optimal form meaning pairs is due to G. Jäger, see [7, 11]. [9] was a first attempt to bring weak optimality together with the notion of *nash equilibria*.
‖ Examples of this type have first been discussed by J. Mattausch [13].

**(iv)** **a)** Hans hat sich ein Rad gekauft.
**b)** Hans hat sich ein Fahrrad gekauft.
**c)** Hans hat sich ein Zweirad gekauft.
**d)** Hans has himself a bicycle bought.

The first two sentences are equivalent but the third is marked. The critical expressions are *Rad*, *Fahrrad* and *Zweirad*. In this context they have all the same meaning, namely *bicycle*. The principle of weak optimality would predict that *Rad* (wheel) is optimal, hence *Fahrrad* (driving–wheel) should tend to have a marked meaning. But both expressions are equivalent. *Fahrrad* and *Zweirad* (two–wheel) are of the same complexity, hence there should be no difference in meaning, but *Zweirad* is marked. In contrast, the following example clearly is in line with weak OT and Horn's principle of division of pragmatic labour:

**(v)** **a)** Hans wischt den Boden mit Wasser/Flüssigkeit.
**b)** Hans mops the floor with water/a liquid.

*Flüssigkeit* (liquid/fluid) clearly indicates that it is not water that Hans uses for mopping the floor.

We observe a difference between a class (A) with example **(iv)** where the hearer has to resolve an ambiguity for interpreting the speaker's utterance, and a class (B) where the critical expressions differ only with respect to their extension. Example **(i)** belongs to class (B), i.e. to examples **(iv)** and **(v)**.

We have seen that we don't get the effects predicted by Bi–OT for class (A). Marked expressions don't show a tendency to go together with the unexpected reading. Our examples which show partial blocking belong all to class (B)¶. Conceptually, this is an important point as the assumption that meaning is highly underspecified is central for Bi–OT. Bi–OT in its naive form makes predictions for both classes.

## 3. Game Theory and Partial Blocking — van Rooy's Principle

We have seen in Section 2 that Bi–OT over–predicts partial blocking if applied too naively. Originally Blutner intended his theory not as a synchronic theory, i.e. as a theory which models the actual reasoning of interlocutors in an utterance situation. Weak optimality was intended to select diachronically stable form–meaning pairs. Soon after emergence of Bi–OT, Game Theory was proposed as a foundational framework [9]. It allows to embed OT within a well understood theory of rational decision. In addition, there has been important work by Prashant Parikh [14, 15] on resolving ambiguities within game theoretic frameworks. For the following discussion we concentrate on van Rooy's paper [16] because he explicitly proposes his theory as a game theoretic explanation of Horn's division of pragmatic labour. Our aim in this section is not so much to show weaknesses of this approach but to show that it applies to different problems.

For simplicity we represent the possible meanings as *attribute–value functions*; i.e. as functions $\mathbf{f} : \mathbf{Feat} \longrightarrow \mathbf{Val}$ from features into values $\{0, 1, -1\}$. Let $m$ be some feature representing some property of objects, $\mathbf{f}$ an attribute–value function, and $e$ an object of type $\mathbf{f}$, $e : \mathbf{f}$. Then $\mathbf{f}(m) = 1$ means that $e$ does have the property $m$; $\mathbf{f}(m) = -1$ means that $e$ does not have the property $m$; and $\mathbf{f}(m) = 0$ means that $e$ may or may not have the property $m$. We denote the set of all attribute–value functions by $\mathbf{Type}$. $\mathbf{f} \in \mathbf{Type}^*$ means that all properties are specified. We call the elements of $\mathbf{Type}^*$ *basic types*. Attribute–value functions are very primitive examples of typed feature structures [8].

¶ There is some work now on anaphora and OT starting with [1]. Examples of class (A) constitute a different type of problem. Hence we restrict our considerations to cases without ambiguities, i.e. class (B). I discussed Mattausch's examples in two previous papers, [3] and [4].

Semantics and pragmatics should tell us what are the optimal forms for the speaker to select and how the hearer interprets them. A *speaker's selection strategy* is a function from meanings into forms; and a *hearer's interpretation strategy* a function from forms into meanings.

Van Rooy observes that if communication shall be successful, i.e. if $H(S(\mathbf{t})) = \mathbf{t}$, then speakers and hearers must coordinate on *separating* strategy pairs $\langle S, H\rangle$, i.e. there must be a subset of forms $\mathcal{F}'$ such that $H \circ S$ maps $\mathcal{F}'$ 1–1 onto $\mathcal{M}$. This implies that it is desirable that speaker's strategies are also *separating*, i.e. that $\mathbf{t} \neq \mathbf{t}'$ implies $S(\mathbf{t}) \neq S(\mathbf{t}')$. Only then can it be guaranteed that every state of affairs can be expressed by language. If the speaker's strategy is not separating, then communication must fail for at least one situation, i.e. there exists $\mathbf{t} \in M$ such that $H(S(\mathbf{t})) \neq \mathbf{t}$. If it is rational for interlocutors to coordinate on strategies where communication is always successful, then the following principle must hold:

**(vi)** Suppose that $F$ is a lighter expression than $F'$, $F > F'$, and that $F'$ can only mean $\mathbf{t}$, but $F$ can mean both. Suppose, moreover, that $\mathbf{t}$ is more salient, or more stereotypical, than $\mathbf{t}'$, $\mathbf{t} > \mathbf{t}'$[+], then speaker and hearer coordinate on strategy pairs $\langle S, H\rangle$ such that $S(\mathbf{t}) = F'$, $S(\mathbf{t}') = F$, $H(F) = \mathbf{t}'$ and $H(F') = \mathbf{t}$.

Van Rooy introduces his principle as a counterexample for Bi–OT. We can represent the situation by the following graph:

**(vii)**

$$\begin{array}{cccc} & \mathbf{t} & & \mathbf{t}' \\ F & \bullet & \longleftarrow & \bullet \\ & \uparrow & & \\ F' & \bullet & & \end{array}$$

It is not difficult to see that van Rooy's principle **(vi)** contradicts Bi–OT and Horn's division of pragmatic labour. Clearly $\langle F, \mathbf{t}\rangle$ is optimal. If we then reduce the graph and eliminate all nodes in the row and column containing $\langle F, \mathbf{t}\rangle$, then no combination remains. Hence, Bi–OT predicts that $F$ denotes $\mathbf{t}$ — and $\mathbf{t}'$ cannot be expressed.

The following examples show that van Rooy's principle is violated in situations of class (B). The claim that interlocutors always coordinate on the separating strategy seems to be incorrect:

**(viii)** **a)** Zwei Amerikaner wurden bei dem Anschlag getötet.
**b)** Mehrere Afrikaner wurden in der S-Bahn angepöbelt.

**a)** Two Americans were in the plot killed.
**b)** Some Africans were in the city train verbally abused.

Without special context these sentences must be understood as:

**a)** Zwei US–Amerikaner wurden bei dem Anschlag getötet.
**b)** Mehrere Schwarzafrikaner wurden in der S-Bahn angepöbelt.

**a)** Two US Americans were in the plot killed.
**b)** Some Black Africans were in the city train verbally abused.

The critical expressions are *Amerikaner* and *Afrikaner*. They have a wider extension than *US–Amerikaner* and *Schwarzafrikaner*. Moreover, they are lighter than the special expressions and the special expressions can only have a special meaning. We can assume that (a) in most cases where Germans talk about inhabitants of the American continent, they talk about US Americans, and (b) Black Africans are more prototypical Africans than North

[+] The first part is cited from [16, Sec. 3.2, p. 13]. The notation is slightly adapted.

Africans; furthermore we can assume that the difference between US–Americans and Non–US Americans and Black Africans and Non–Black Africans is relevant. If we naively apply van Rooy's principle, then we should expect a tendency for *Amerikaner* to denote Non–US Americans, for *Afrikaner* to denote North–Africans, etc. But we observe the opposite effect.

It is not confined to examples where we classify people according to their nationality:

**(ix)** **a)** Hans macht Urlaub in *Amerika*.
**b)** Hans fährt seinen *Wagen* in die Garage.
**c)** Hans makes holidays in *America*.
**d)** Hans drives his *car* into the garage.

The first example must be understood as meaning that Hans makes holidays in the USA, not e.g. in Chile. *Wagen* can have a very wide meaning including both a car and a hand cart. The lighter, more general expression has always the tendency to denote the normal case. What if van Rooy's principle could be applied to these examples? It would predict the contrary effect. Van Rooy's principle is violated in class (B) — if applied too naively, of course. By *applying naively* I mean: applying without checking the preconditions. There are two reasons for why van Rooy's models cannot be used for class (B). He has to assume that the meaning of some forms is underspecified. Then, he has to start with non–separating signalling systems, and try to show that they develop into separating ones. This implies that the models cannot be applied if:

i. Forms have unique meanings.
ii. Languages are separating.

This is the situation we find in examples of class (B). We can always assume that natural language is fine–grained enough to express every state of affairs, i.e. we can assume that natural language is separating. Hence, the central problem with partial blocking phenomena is to explain how there can be shifts in meaning for signalling systems that are (a) separating and (b) unambiguous. If this is true, then partial blocking poses a type of problem which is sharply differentiated from the problems approached by van Rooy or Parikh.

## 4. Associative Learning and Partial Blocking

For the introductory example **(i)** it has to be explained why *kill* tends to denote a *typical* killing event whereas *cause to die* tends to denote an *untypical* killing event. I want to show that partial blocking can be explained as an effect of *associative learning* and speaker's preferences. It emerges as the result of a process which divides into the following stages: (1) In the initial stage all killing events are direct killing events. The speaker will always use *kill* to denote these events. (2) Interpreters will learn that *kill* is always connected with direct killing. They *associate kill* with direct killing. (3) The speaker will learn that hearers associate *kill* with direct killing. (4) If then an exceptional event occurs where the killing is an indirect killing, the speaker has to avoid misleading associations and use a different form. In this case it is the more complex form *cause to die*. (5) The hearer will then learn that *cause to die* is always connected to an untypical killing. By *associative learning* we mean the learning process in (2), (3) and (5). For the hearer I assume that the following principle holds:

(H) In every actual instance where the form $F$ is used for classifying events or objects it turns out that the classified event or object is at least of type $\mathtt{t}$, then the hearer learns to associate $F$ with $\mathtt{t}$, i.e. he learns to interpret $F$ as $\mathtt{t}$.

A similar principle is assumed for the speaker to explain step (3):

(S) In every actual instance where the form $F$ is used for classifying events or objects it turns out that the hearer interprets $F$ as $\mathbf{t}$, then the speaker learns that he can use $F$ for expressing $\mathbf{t}$.

It is not only word meaning that is involved:

**(x)** The dress is pink/pale red/pale red but not pink.

All three phrases, *pink*, *pale red*, and *pale red but not pink*, are forms which the speaker can choose. The forms $F$ may even be lengthy descriptions of a situation.

A formal model must contain the following elements: (1) A set of possible meanings for words and phrases. (2) A representation for the semantics of a given language $NL$. (3) A representation for the speaker's preferences on forms. We do this by adding a pre–order $\preceq$ on $NL$, where $F \prec F'$ means that $F$ is less marked than $F'$.

Less obvious from the previous discussion is that we will need also: (4) A representation for the speaker's knowledge about the object or event he wants to classify. (5) A representation for the speaker's intentions on how to classify an object or event.

We consider settings of the following form: There is an object or event $e$ and the speaker wants to classify it as being of a certain type $\mathbf{f}'$. Maybe he knows more about the object, maybe he knows that it is in fact of a more special type $\mathbf{f}$. But all he wants to communicate is that it is of type $\mathbf{f}'$. He has to choose a form $F$ such that the hearer can conclude that the object or event $e$ is of type $\mathbf{f}'$. This explains why we need a representation for speaker's knowledge and intentions. We represent them by attribute–value functions.

These elements form the *static* part of our model. What does change diachronically? (6) The types of objects and events which actually occur. We represent the actual occurrences of objects and events during a period $\alpha$ by a set $E_\alpha$. (7) The hearer's interpretation of forms. We represent it by a function $H$ from forms into meanings. (8) The speaker's choice of forms. We represent it by a function $S : \langle \mathbf{f}, \mathbf{f}' \rangle \mapsto F \in NL$, i.e. a function which maps pairs of attribute–value functions which represent his knowledge ($\mathbf{f}$) and intentions ($\mathbf{f}'$) into forms. We assume throughout that the speaker is truthful and sincere; this means especially that $\mathbf{f}'$ represents not more information than $\mathbf{f}$. The functions $S$ and $H$ are the counterparts of the speaker's and hearer's strategies in game–theoretic approaches.

We noted in the last section that the central problem with partial blocking phenomena is to explain how there can be shifts in meaning for signalling systems that are (a) separating and (b) unambiguous. We assume that in the initial situation choice and interpretation of language is governed by its (unambiguous) semantics. Let us denote the meaning of a form $F$ by $[F]$, and assume that for every meaning $\mathbf{f}$ there is at least one form $F$ such that $[F] = \mathbf{f}$. The speaker should select the optimal form:

$$S^0(\mathbf{f}, \mathbf{f}') := \min\{F \in NL \mid \mathbf{f} \leq H^0(F) \leq \mathbf{f}'\}.$$

The hearer's initial interpretation should simply follow the rules of pure semantics; i.e. $H^0(F) = [F]$. The definitions imply that

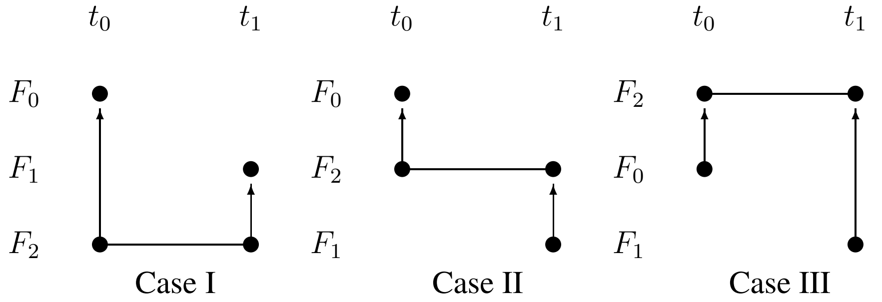
$$\mathbf{f} \leq H^0(S^0(\mathbf{f}, \mathbf{f}')) \leq \mathbf{f}', \tag{4.1}$$

i.e. the speaker will always have success. In addition we assume that the speaker does classify entities correctly.

### *The Situation with two Basic Types*

We look at a special case: the situation for one feature with two values. The examples considered so far are of this type, at least after some simplification of the scenarios. E.g. in **(i)** the question was whether the killing is *direct* or not. Hence we can assume one feature *direct* with possible values $-1$ and $1$ for *not direct* and *direct*. In **(v)** the question was whether it is *water* or not that Hans uses for mopping the floor.

If we consider a situation with two basic types $\mathbf{t}_0$ and $\mathbf{t}_1$, then there are only three forms $F_0, F_1, F_2$ the speaker has to consider for making his choice. Without loss of generality we can assume that $[F_0] = \mathbf{t}_0$, $[F_1] = \mathbf{t}_1$ and $[F_2] = \mathbf{t}_0 \vee \mathbf{t}_1$. Hence, $F_2$ always denotes the form with the wider meaning. We can further assume that in general $F_0$ is preferred over $F_1$. Hence, we arrive at the following classification of all situations with two basic types:




The topmost form is the most preferred one, the lowest the least preferred. The vertical arrow indicates the speaker's preferences. The horizontal line means that the respective form has an extension which comprises the meaning of both types $\mathbf{t}_0$ and $\mathbf{t}_1$. Examples are: Case I *father, mother, one of the parents* ($F_0 \prec F_1 \prec F_2$); Case II *water, liquid, alcoholic essence* ($F_0 \prec F_2 \prec F_1$); Case III American, North American, Latin American ($F_2 \prec F_0 \prec F_1$).

Hence, we see that **(v)** is a Case II example. What about *kill–and–cause–to–die* **(i)**? We may assume that the relevant forms are $F_2$ = *killed*, $F_0$ = *directly killed*, and $F_1$ = *indirectly killed*, hence it belongs to class III. For the classification we considered only the most economic forms for each type. We add $F_3$ = *caused to die* and assume for simplicity that $F_2 \prec F_3 \prec F_0 \prec F_1$. This is a sub–case of Case III. How can we explain the observed differentiations in meaning between $F_2$ and $F_3$? We claimed that we can see it as the result of a diachronic learning process. This process stretches over a sequence of *(synchronic) stages*. We have to describe how selection and interpretation strategies change from stage to stage. What is a synchronic *stage*? It is a triple $\mathrm{Syn}_i = \langle E^i, S^i, H^i \rangle$ where

$$E^i \subseteq E \times \mathbf{Type} \times \mathbf{Type} \ \ \& \ \ \langle e, \mathbf{f}, \mathbf{f}' \rangle \in E^i \Rightarrow (e : \mathbf{f} \,\&\, \mathbf{f} \leq \mathbf{f}'). \tag{4.2}$$

This means that every synchronic stage is characterised by (1) the set of utterance situations which comprises a classified entity $e$, the speaker's knowledge $\mathbf{f}$ about $e$, and his intentions to classify $e$ as $\mathbf{f}'$; (2) the speaker's selection strategy; and (3) the hearer's interpretation strategy.

We repeat the informal description of the principles governing the hearer's learning in each stage:

(H) In every actual instance where the form $F$ is used for classifying events or objects it turns out that the classified event or object is at least of type $\mathbf{f}$, then the hearer learns to associate $F$ with $\mathbf{f}$, i.e. he learns to interpret $F$ as $\mathbf{f}$.

The following definition contains the idea of the paper in a nutshell. Assume we are in stage $\mathrm{Syn}_n = \langle E^n, S^n, H^n \rangle$. How do the new selection and interpretation strategies in the next stage $Syn_{n+1}$ look like?

$$H^{n+1}(F) \ := \min\{\mathbf{f} \in \mathbf{Type} \mid \mathbf{f} \leq H^n(F) \wedge \|F\|_n \subseteq [\![\mathbf{f}]\!]_n\} \tag{4.3}$$

$$S^{n+1}(\mathbf{f}, \mathbf{f}') := \min\{F \in NL \mid \mathbf{f} \leq H^{n+1}(F) \leq \mathbf{f}'\}. \tag{4.4}$$

Where $[\![\mathbf{f}]\!]_n$ denotes the *extension* of $\mathbf{f}$ in $E^n$, i.e. $[\![\mathbf{f}]\!]_n := \{e \in E^n \mid e : \mathbf{f}\}$; $\|F\|_n$ is the set of all entities where the speaker has in fact used $F$ to classify them, i.e. $\|F\|_n := \{e \in E \mid \exists \mathbf{f}, \mathbf{f}' :$

$\langle e, \mathbf{f}, \mathbf{f}' \rangle \in E^n \wedge S^n(\mathbf{f}, \mathbf{f}') = F\}$. $H^{n+1}$ and $S^{n+1}$ describe both the hearer's and the speaker's learning. The hearer's learning precedes the speaker's, but we put both processes together in one stage. This learning should take place only with respect to actually used forms. If a form is never used, then the hearer can associate no restricted information with this form. Hence, we have to check which forms are used in each stage. We collect them in a set $NL_{n+1}$:

$$NL_{n+1} := \{F \in NL_n \mid \exists (e, \mathbf{f}, \mathbf{f}') \in E^n \; S^n(\mathbf{f}, \mathbf{f}') = F\} \tag{4.5}$$

If learning takes place with respect to $NL_{n+1}$ only, then we have to restrict the definition of $H^{n+1}$ in (4.3) to this set. The actual selection and interpretation functions $H^{n+1}$ are defined by:

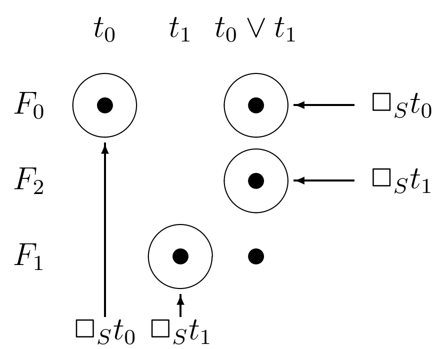
$$H^{n+1}(F) := \begin{cases} H^n(F) \text{ for } F \notin NL_{n+1} \\ H_*^{n+1}(F) \text{ else} \end{cases}, \tag{4.6}$$

where $H_*^{n+1}$ is the function defined in (4.3). The diachronic model $(\mathrm{Syn}_i)_{i=0,\ldots,n}$ is totally determined by semantics and speaker's preferences on forms.

Let us apply this model to the *Kill–and–cause–to–die* Example **(i)**! The observed interpretations emerge as the result of a process involving two stages: (1) In the initial stage all killing events are direct killing events, i.e. in the first stage $\mathrm{Syn}_0$ there are only events $e$ which represent direct killings. The speaker will always use *kill* to denote these events. Hence, interpreters will learn that kill is always connected with direct killing. They *associate kill* with direct killing. The relevant types are $\mathbf{t}_0$ = *direct killing* and $\mathbf{t}_1$ = *indirect killing*. Hence, we find $H^1(F_2) = \mathbf{t}_0$ and therefore the speaker will learn that hearers associate *kill* with direct killing. We observe further that the situation turns from a class III example into a class II example with $F_2 \prec F_3 \prec F_1$.

(2) In the second stage $\mathrm{Syn}_1$ the speaker encounters an instance $e'$ of an indirect killing. He has to avoid misleading associations and use the more complex form *cause to die*. We find that $S^1(\mathbf{t}_1, \mathbf{t}_0 \vee \mathbf{t}_1) = \min\{F \in NL \mid \mathbf{t}_1 \leq H^1(F) \leq \mathbf{t}_0 \vee \mathbf{t}_1\} = F_3$. He cannot select $F_2$ because $\mathbf{t}_1 \not\leq H^1(F_2)$. $F_3 \notin NL_1$, hence $H^1(F_3) = \mathbf{t}_0 \vee \mathbf{t}_1$. If we assume that the speaker always knows whether it was a direct or an indirect killing, then the hearer will learn that *cause to die* is always connected to an indirect killing $\mathbf{t}_1$; hence $H^2(F_3) = \mathbf{t}_1$. This in turn can be exploited by the speaker, and he will start to use *cause to die* for expressing $\mathbf{t}_1$.

Let's turn to Example **(v)**. We provide a graphical solution. The first row in the graph represents the speaker's possible intentions on how to classify an object. $\Box_S$ is to be read as *the speaker knows that…* Hence, $\Box_S \mathbf{t}_0$ means that the speaker knows that the entity he classifies is of type $\mathbf{t}_0$. The circles around bullets are to indicate that these form–meaning pairs are optimal according to his preferences. The arrows from $\Box_S t_i$ indicate that this optimality depends on the speaker's knowledge. The situation for Case II examples looks as follows:



We can see that the speaker will use the general form $F_2$ only if he knows that the entity $e$ has to be classified as being of type $\mathtt{t}_1$. Hence, as a matter of fact, if the hearer knows that the speaker knows the type of $e$, he can safely infer from an utterance of $F_2$ that the entity is of type $\mathtt{t}_1$. This explains why *Flüssigkeit* in **(v)** is interpreted as meaning *not water*.

So even a first survey shows how associative learning can lead to stronger interpretations and differentiations of meaning. Moreover, the survey provides us with a classification of utterance situations.


[1] D. Beaver (2000): *The Optimization of Discourse;* ms. Stanford; to appear in *Linguistics and Philosophy*.

[2] D. Beaver, H. Lee (2003): *Input–Output Mismatches in OT;* To appear in: R. Blutner, H. Zeevat (eds.): *Optimality Theory and Pragmatics*. Palgrave/Macmillan..

[3] A. Benz (2001): *Towards a Framework for Bidirectional Optimality Theory in Dynamic Contexts;* ms., Humboldt Universität Berlin. Available as ROA 465-0901 and from http://www.anton-benz.de.

[4] A. Benz (2003): *On Coordinating Interpretations - Optimality and Rational Interaction ;* To appear in P. Kühnlein, H. Rieser, H. Zeevat (eds.): *Perspectives on Dialogue in the New Millennium*; preliminary paper available from http://www.anton-benz.de.

[5] R. Blutner (1998): *Lexical Pragmatics;* Journal of Semantics 15, pp. 115–162.

[6] R. Blutner (2000): *Some Aspects of Optimality in Natural Language Interpretation;* In: Helen de Hoop & Henriette de Swart (eds.) *Papers on Optimality Theoretic Semantics*. Utrecht Institute of Linguistics OTS, December 1999, pp 1-21. Also: *Journal of Semantics* 17, pp. 189-216.

[7] R. Blutner, G. Jäger (2000): *Against Lexical Decomposition in Syntax;* In A.Z. Wyner (ed.): *Proceedings of the Fifteenth Annual Conference*, IATL 7, University of Haifa, pp. 113-137 *Proceedings of IATL 15*, University of Haifa.

[8] B. Carpenter (1992): *The Logic of Typed Feature Structures;* Cambridge University Press, Cambridge.

[9] P. Dekker, R. v. Rooy (2000): *Bi–Directional Optimality Theory: An Application of Game Theory; Journal of Semantics* 17, pp. 217–242.

[10] L. Horn (1984): *Towards a new taxonomy of pragmatic inference: Q–based and R–based implicature;* In: D. Schiffrin (ed.): *Meaning, Form, and Use in Context: Linguistic Applications*, Georgetown University Press, Washington, pp. 11–42.

[11] G. Jäger (September 2000): *Some Notes on the Formal Properties of Bidirectional Optimality Theory;* ms, ZAS Berlin.

[12] D. Lewis (1969): *Convention;* Harvard University Press, Cambridge.

[13] J. Mattausch (November 2000): *On Optimization in Discourse Generation;* master thesis, Universiteit van Amsterdam.

[14] P. Parikh (1990): *Situations, Games and Ambiguity;* In R. Cooper, K. Kukai, J. Perry: *Situation Theory and its Applications I*, CSLI Publications, Stanford.

[15] P. Parikh (2001): *The Use of Language;* CSLI Publications, Stanford.

[16] R. van Rooy (2002): *Signalling Games select Horn Strategies;* ms Universiteit van Amsterdam; to appear in Linguistics & Philosophy.