Bayesian Interpretation and Optimality Theory

Henk Zeevat
ILLC, University of Amsterdam
henk.zeevat@uva.nl

1 Introduction

Since a couple of years I have been defending that pragmatics should be integrated with semantics and syntax by combining the proper pragmatical constraints with a strongest constraint FAITH that grosso modo says: the candidate interpretation should have the utterance as an optimal realisation, using production optimality theory, i.e. a constraint system that defines optimal realisations by combining OT syntax with OT phonology for a semantic input in a context of utterance (Zeevat, 2001, 2007, 2009).

This way of proceeding gives good results in pragmatics, but it raises the question of how the brain is able to invert production OT. It is clear that the same or another system of constraints in an inverse competition will not give the inversion, unless the constraint system is special. The algorithms for processing OT do not help therefore. And if it cannot work in this way, one should have an account of how the brain manages to do the inversion of the relation as effortlessly as it seems to do it.

There is a growing body of evidence that the production system is activated in understanding. (Galantucci et al., 2006): gives evidence of mirror neurons firing both in language production and in understanding. This means that the assumption that the brain assigns a role to production in understanding is reasonable.

Traditional grammar is concerned with investigating the relation between forms and meanings for particular natural languages. Natural languages tend to be massively ambiguous due to lexical ambiguities, ambiguous constructions and ambiguous context integration processes. This means that grammar is not useful for explaining the problem known as parity (Liberman and Mattingly, 1985), coordination on a meaning (Clark, 1996) or intention recognition (Grice, 1957), the explanation of how a speaker and a hearer can converge on the same meaning given the production of an utterance by the speaker. Traditional grammar predicts that convergence is a a rare though possible event.

Instead, the hearer should be trying to find the meaning E for which $p(E|O)$ (the probability that the meaning is E given the utterance O and the context, here and elsewhere the context will not appear in the notation) is maximal (stochastic interpretation) and in order to avoid misunderstanding the speaker should be ensuring that O is such that $p(E|O)$ is maximal for his intended meaning E (self-monitoring). The problem is however that p(E—O) is not directly accessible. There is no theory that predicts $p(E|O)$ and given that the number of utterances is in principle infinite, direct counting is impossible and estimation by some formula is the only option. Especially in combination with the idea that in arriving at an interpretation involves many levels, this means that a quite substantial error is unavoidable.

In principle, stochastic interpretation and self-monitoring could employ a model of $p(E|O)$ for finding the best interpretation and for checking that the hearer will get E right. But it seems that this is not the best option. In the capacity of expressing meanings in utterances, the brain has a way to estimate $p(O|E)$. Moreover, in its capacity of perception, the brain must have developed a powerful and accurate estimation technique for comparing the a priori probability of the percepta. These can be combined in an emulation of Bayesian interpretation which will outperform using a directly estimated model of $p(E|O)$, because the estimations are more accurate.

Bayesian interpretation is finding E such that $p(E|O)$ is maximal by a different route: finding E such that $p(E)p(O|E)$ is maximal. It follows from Bayes's theorem that $p(E|O)$ is maximal for an E for which $p(E)p(O|E)$ is maximal. If the brain can use its capacity for producing utterances to estimate $p(O|E)$ accurately and its model of $p(E)$ for perception is similarly accurate it must be able to outperfor sequential data-derived estimates of $p(E|O)$.

This paper argues that the brain runs an emulation of Bayesian interpretation in which the probability of the message in the context is maximised within the space of the primed possible interpretations (it will be made plausible that these are always maxima for the probability of the message, given the way priming works) by using production simulation to hit the maxima for the probability of the message causing the utterance. The two processes together should give the message E that gives a maximum for $p(E)p(O|E)$ for the given utterance O.

Production OT together with OT learning can be seen as a sophisticated estimation technique for $p(O|E)$. And pragmatics — as implemented by priming— comes out as a technique for finding maxima in $p(E)$. Pragmatics in this conception is dependent on extra-linguistic resources, i.e. on the resources that are needed in perception and more generally in the explanation of natural phenomena and the behaviour of other agents. These resources estimate the plausibility of perceived situations and of hypotheses in explanations and are acquired by learning from experience.

A good OT model of production is able to do a reliable estimate of $p(O|E)$: the estimation can be computed by running production with the system many times with a noise factor. It can deal with ill-formed input by making the noise factor high. If a general model of plausibility would be available (currently, it is not), estimation would also be possible for $p(E)$. The estimates so obtained could then be used directly to obtain the maximum of $p(E|O)$ for a given O using dynamic programming. But this direct computation cannot use O to restrict the space of hypotheses: it needs to start with an arbitrary interpretation and then use the dynamics to make E more plausible and a better explanation of

O.

Natural dialogue abounds in corrections, self-corrections and feed-back loops (Clark (1996)). This is enough to argue that communication is an uncertain business. It is not for nothing that in current computational linguistics, the ambiguity problem is the central problem. Communication therefore must be seen as a risky business with interpretation the shakiest part. It follows that the brain must recruit any resources it can lay its hands on to improve its chances of getting interpretation right. So Bayesian interpretation would have emerged as a way of improving interpretational success rates.

To see this, let us assume with most of the literature that the speech signal passes through a series of intermediate representations before it is represented as an update of the hearer's information: a phonetic represention, a phonological representation, a string of words, a string of morphemes, a labeled tree bottoming out in morphemes, a tree of concepts, a logical structure and a pragmatic structure. The direct method would have to come up with models of $p(R_{n+1}|R_1, ..., R_n, C)$ for each higher representation R_{n+1}. But there is no way of doing a direct empirical determination of the probabilities for any of the cases: the size of the representations is arbitrarily large and the same holds for the context. That means that one needs to use finite approximations and formulas and that error is to be expected. While this is the only rational procedure, it runs into the problem that the total map from speech signal to context update has to be found by composition and that the success rate will be s^n where n is the number of representational levels and s the mean success rate for any of the probability models. For $n = 7$ as above and s an optimistic 0.9, this gives a compound success rate below 0.5.

So it seems plausible that estimating $p(E|O)$ directly on the basis of learning data (apparently correctly understood utterances of others and utterances of oneself that were apparently correctly understood) does not by itself give a realistic result. The brain cannot use this method since it cannot train on as many data as in typical training corpora and it is not capable of the sophisticated statistics used in this kind of stochastic processing either.

In the model proposed, the direct method is only the starting point of the interpretation process: it supplies a set of hypotheses by a priming process that combines finding interpretations on the basis of the utterance with pragmatic optimisation. It can be assumed that the hypotheses are ordered by pragmatics: the most probable E is most activated. Simulation is used to converge on the most activated E that also explains O. Now suppose E is found by the direct method and is a maximum for $p(E|O)$. It is then also a maximum for $P(E)p(O|E)$. On the assumption that $p(E)$ is reliably estimated by priming and that $p(O|E)$ is reliably estimated by simulation, E is then retained as athe winner with a high probability. If E is not correct, there is an E' with a higher plausibility or with a higher $p(O|E')$. If E' is more plausible, it will be considered before E by priming and E has no chance of putting E' out of action. If $p(O|E') > p(O|E)$ and E' is among the hypotheses, E' will be preferred in the scan of all the hyoptheses found by priming. So assuming that $p(E)$ and

$p(O|E)$ are reliably estimated, the correct predictions of the direct model are retained and the incorrect predictions are corrected, if priming finds the correct hyopthesis as a possible one.

The fact that the direct route from signal to interpretation is not very reliable and that combination with simulation is much more so makes it plausible that evolution has recruited simulation for better language understanding. It is quite plausible that much the same applies to many other kinds of perception/explanation where simulation is possible, e.g. in motivational explanations of other people's behaviour[1] . If one believes in a cognitive derivation of our notions of causality from the capacity to change the world by our planned action, it also becomes plausible that simulation plays a role in those areas. Simulation is then primarily a technique for improving the quality of observation and explanation and should not be equated with understanding itself, as in theories of analysis by synthesis.

Section 2 tries to give a direct account of how the brain can go about emulating Bayesian interpretation. Section 3 describes how a particular optimality theoretic account of pragmatics can be seen as the priming bias in finding possible interpretations. Section 4 relates production OT to the probability $p(O|E)$. Section 6 discusses the complement of production simulation: understanding simulation in production and shows that it is just as necessary for success of communication. It moreover is needed as an addition to production OT to capture the facts of language use.

2 Emulating Bayesian Interpretation

This section tries to give some more body and motivation to the model.

Following Grice (1957), interpretation should be seen as having a natural endpoint, in which the intention of the speaker is grasped. This can be n intention to inform, to query, to make a request, to promise something under various modalities. The recognition involves convergence on the content: on what information is given, what is asked, what is requested or promised, which in turn involves the disambiguation of the syntactic structure, the words used, and the correct resolution of the connections between the concepts among themselves and between the concepts and the linguistic and non-linguistic context.

There are a number of levels at which we are conscious of the interpretation process.

[1] E.g. Kilner et al. (2007) is an attempt to use a Bayesian interpretation scheme to perceiving the intentions behind gestures.

(1) the phonemes
the phonological words
the concepts expressed
the referents if any
the semantic relations
the message
the new context
the speaker's intention

The transitions between these representations can be described as arrows between a higher and a lower representation.

We assume that all the arrows are essentially association arrows: the activation of the higher level activates the lower level and inversely. It is association strength that determines which other representation is most activated in both directions.

But there is one further assumption as well. In the activation of a higher level from a lower level, also the inverse arrow is activated and influences the activation levels on the lower level. If the original input on the higher level is obtained, this has no effect. If it is however different or difficult to obtain, this has an inhibiting effect on the lower level representation and can lead to other representations on the lower level to be the most activated.

And the other way around: from lower level to higher level it works just the same.

The inverse activation and the inhibitory effect on certain interpretations implements the simulation test on interpretations. Not always, since the utterance may be ill-formed for different reasons. But it is hard for an interpretation to win if there is another interpretation for which the utterance is perfect, unless that interpretation has a very low prior probability. In such a case, the hearer is obliged to offer an explicit correction of the speaker's utterance. Understanding an imperfect utterance however always comes with an correction which the interpreter can offer to the speaker as a means to help the speaker with his language skills or to obtan feedback.

In the other direction, the inverse arrow gives self-monitoring. The speaker can see in this process whether the hearer will understand him as he intends and make amends during and after generation for perceived problems.

Gestalt Effect

This is the effect in perception that a perception is only finished when it has become the perception of something. It is called "the unity of apperception" by Kant. Grice isolates exactly this aspect of language interpretation (and more) in his analysis of non-natural meaning. One can reformulate the insight as: an utterance recognition is its recognition as an attempt by another person to bring about something by means of producing the utterance based on the recognition of the intention to bring about that something.

This can be rephrased for our purposes as: the interpretation of an utterance is incomplete unless it contains the attribution to the producer of her intention in producing the utterance: the reason for its producer to produce it. The intention should be formulated in terms of the effect. It should also be rich enough to explain all aspects of the utterance.

Linking

An important part of interpretation is the linking of concepts. Nearly all concepts are essentially incomplete[2]. The concept of tearing up in “he tore up the paper“ is incomplete before it is linked to an agent that does the tearing up and to an object that gets torn up. The other basic concepts in the example sentence have the same incompleteness: *he* is incomplete in evoking a highly activated male person before this person is identified in the context, the past tense before it is connected to a reference time, *the paper* before a similar identification with an object in the context. All such links have to be established before an interpretation is complete.

Semantic Memory

A linked concept (with the concepts to which it links) can be an instance of a frequent pattern or it can be rare. It seems one can use frequent patterns as reinforcers of links and concepts and that frequency can be learnt from experience. A fully linked sets of concepts can be preferred over another by the existence of more reinforcing patterns. The knowledge resource for this kind of processing can be equated with semantic memory or conceptual knowledge.

Representations

One needs to represent the utterance at a number of levels. As a sound signal, as a phonological structure, as a structure of lexemes, a structure of concepts and as a new context model that integrates the new information achieved in the interpretation. The structure of lexemes can be identified with a syntactic structure —if this is necessary— the structure of concepts with a logical form. It should however be clear that a new context in which the interpretation is complete contains all the information from which syntactic relations and logical relations can be reconstructed. It is quite tempting to say that building all the conceptual links in fact and completing the interpretation in fact achieves the new context and constructs the logical and the syntactic relations and thereby is also the logical and syntactic structure.

[2] Frege makes the assumption that names are complete. But many people nowadays would dispute that and give names the status of pronouns with a special condition on what the pronoun should be resolved to, an individual that is named so and so. If that is so, there is no obvious exception to the claim that all concepts are incomplete. Perhaps one word utterances like ”Ouch” or ”Help” are the best candidates of concepts expressed that are complete. It still does not seem so: the person who is in pain and the cause of the pain should be recovered if possible. Also who needs to help who with what. The only properly complete object is a fully linked coherent set of concepts, corresponding with Frege’s type of truth-value bearers.

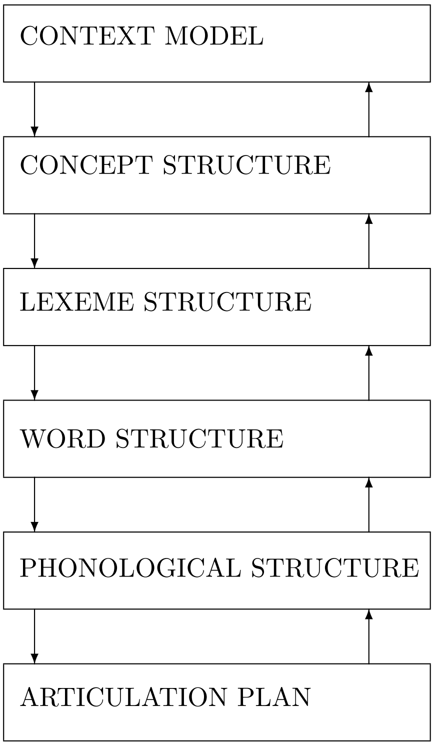

Figure 1. A hierarchy of representations and their connections.

Bidirectional inter-representation arrows

The different representations evoke each other in both directions by some mechanism. It is natural to assume that the mechanisms for each arrow are always switched on so that the achievement of a representation based on a lower representation starts trying to construct a representation on the lower level again (or inversely). If the source representation is reached again this reinforces the new representation, if a match is not obtained this inhibits the new representation.

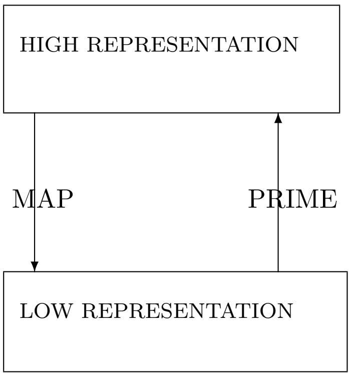

Figure 2. The configuration of two levels of representation and the two arrows. Both arrows can be inhibited by the other arrow not given back their input.

Context model

The context model can be equated with a set of activated concepts linked to other concepts in the set. Some of these concepts happen to be concepts of individual objects.

An interpretation is just a new context model: the old set of linked concpets with possibly some parts deleted and the new concepts added with appropriate links to the old ones. Since there is not much semantics and pragmatics without links to the context, so the end-point of interpretation is best seen as a new context, as in DRT Kamp (1981) or other dynamic models of NL interpretation.

Given all of the above, the model proposed in this paper is just the assumption that from lower to higher representation the mechanism is priming, with biases due to recency, frequency and relevance and that from higher to lower representation the mechanism is a process of representation transformation.

An intended new context model is mapped to a set of linked concepts that represent the contextual amendment that the speaker envisages given the rest of the context, the linked concepts to a structure of lexemes needed for their expression, the lexeme structure to a sequence of words or a tree with words as leaves, the tree into a phonological structure and the phonological structure to a motor program for the articulatory organs. There may be some freedom (most obviously in lexical choice) but it is fairly clear what the lower representation must be like in terms of the higher representation and no further resources are required. The maps can be simulated by computer programs that just look at the higher representation and change it to the lower representation. In fact, if the higher structures are directly connected to what is needed for the construction of the lower representation, the computer program only needs to look at the higher representation and what it is connected to. For this purpose, concepts should be connected to lexemes, lexemes to morphemes and words to their phonological structure and phonemes to articulatory gestures.

The process emulates Bayesian interpretation under three assumptions that are approximately true, though it is certainly possible to find counterexamples. But all that matters for a good emulation is that the counterexamples are not too frequent.

Good heuristics:
The upward links activate the interpretation E for which $p(E|O)$ is maximal for the given O more than any other interpretation E' that the downwards links would also map to O.

The assumption entails a property of the upwards arrows and a property of the downwards arrows:

Syntax:
For all E, the function $f_E(O) = p(E|O)$ has few peaks.

Priming adequacy:
the priming process gives neither too few nor too many interpretations.

One peak only makes *good heuristics* trivial, any additional peak increases the risk of missing the most probable interpretation. Too many interpretations also increases that risk while too few interpretations creates the risk of not finding the best interpretation at all.

The second assumption is the following

Good priming:

The bias of the priming mechanism is such that if E is more activated by it from input O than E', then $p(E') < p(E)$.

It is not difficult to violate *goodheuristics*. Try understanding *each other* in the first example of (2) in the way it is understood in the second example: this seems impossible. A production of the first example to express that interpretation would be interpreted in competition with the natural interpretation and the intention would not be recognisable. If the production is syntactically allowed and can be produced for the unnatural interpretation, it would be an occasion in which good heuristics fails. (The proper description should be that the production is not allowed for the unnatural interpretation, though not for syntactic reasons.)

(2) Katja and Henk were surprised that the editors rejected each other's papers.
Katja and Henk were surprised that the journal rejected each other's papers.

Priming bias can work against the intended interpretation. This is often the case when one is misheard or misunderstood. The assumption of *good priming* merely says that there would be more such incidents with other heuristic systems.

If *good priming* holds, the upwards arrows prefer maxima of $p(E)$ that are consistent with the signal. By *good heuristics*, one of those is the interpretation E such that $p(E|O)$ and $p(E)p(O|E)$ is maximal. It is selected on the path that priming selects through interpretation space by being the first one for which $p(O|E)$ reaches a serious value (about $\frac{1}{n}$ for a small n by the assumption of *Syntax*): this local maximum is recognised because O can be produced from E in a probability peak.

In terms of the hill-walking metaphor, priming defines a path over the hills in which $p(E)$ always reaches the maximal value consistent with O among the Es which are not yet discarded. The path will eventually lead to the E for which $p(E)p(O|E)$ is maximal. By *good heuristics*, that can be equated with the first E for which $p(O|E)$ reaches a peak value. $p(E)$ goes down on the path, so it is also the lowest value for $p(E)$ that has been considered so far.

The process does not crash on ungrammatical, badly pronounced or otherwise defective input. It can revert to a maximum for $p(E)p(O|E)$ encountered earlier on (the simulation is quite unlikely but achievable by assuming sufficient noise). (It can get too bad however. In that case, no E is better than any other.)

Priming bias can however also be responsible for misunderstandings. (3a) has (in normal contexts) the interpretation in which Bill's pushing caused John to fall. So language production structurally allows unmarked causes in the second clause. But when one tries that principle as in the second example, Mary's smile is more probably interpreted as Mary's reaction to John falling. The causal interpretation is however quite possible, witness the c-example. Saying (b) intending the interpretation of (c) will lead to a misunderstanding, without there being anything wrong with the interpretation.

(3) (a) John fell. Bill pushed him.
(b) John fell. Mary smiled at him.
(c) John fell. Because Mary smiled at him.

One further assumption eliminates this category of misunderstandings and so makes communication better in the sense that the speaker's intention will be recognised with a greater probability.

Monitoring:
The speaker chose and executed his signal O for E in such a way that $p(E|O)$ is maximal among the choices for O that give peaks for $p(O|E)$.

This assumption is not needed for the emulation of Bayesian interpretation. But if *monitoring* holds with *good priming* and *good heuristics*, communication cannot go wrong anymore: the hearer just finds the interpretation that the speaker intended. Unfortunately speakers and hearers are fallible, life is noisy and it is not unlikely that *good heuristics* and *good priming* will fail at various little used spots in the language or that *monitoring* does not lead to better results. Section 6 will discuss monitoring.

3 Pragmatics as $p(E)$ maximation

It is not usual to say that pragmatics is about the maximation of the probability of the message, but this is inevitable if pragmatics is seen as an explanation of the utterance. The explanation would be the intention attributed to the speaker and the best explanation is the most likely one. Hobbs et al. (1990) is a good example of an approach of this kind, another one is Hamm and van Lambalgen (2005), but interpretation as explanation is a common approach in Artificial Intelligence. Other formal approaches to pragmatics start from model minimalisation, in terms of the number of objects or in terms of the size of certain predicates (Schulz (2007)), but it will be shown that this can also be seen as probability maximation[3].

The OT approach to pragmatics of Zeevat (2009)[4]—inspired by Blutner and Jaeger's formalisation of presupposition in OT (see Blutner (2000) for a discussion) and Mattausch's analysis of optional rhetorical marking— can also be seen as finding the best explanation of the utterance. Natural language pragmatics in this view consists of a system of three ordered constraints (ordered

[3] Informal approaches such as Grice (1975); Sperber and Wilson (8695); Van Rooy (2003); Levinson (2000); Horn (1984) or Blutner (2000) can be brought under this umbrella. These proposals describe ideal behaviour of the speaker that the hearer has to assume as the standard case, or ideal behaviour of the hearer that the speaker has to accommodate to or face the consequence of being misunderstood or constraints on the behaviour of the speaker (that can figure directly in the explanation) and on the behaviour of the hearer (for which the speaker has to make allowances). These proposals would then prefer certain explanations over others, nl. the ones where the speaker complies with the prescribed behavioural constraints or makes allowances for what the hearer is predicted to do.

[4] The argument that this system suffices for Gricean implicatures, presupposition and discourse structure is made in that paper. The application to discourse structure is further elaborated in Zeevat (2007)

as indicated) that can all three be seen as selecting better explanations for utterances.

1. PLAUSIBLE

the probability that the explanation holds given the explainer's knowledge exceeds that of its competitors.

2. *NEW

If one explanation is contained in another, the larger explanation can be eliminated. If two explanations differ only in connecting a new object to an old object or not, the non-connecting explanation is eliminated.

3. RELEVANCE

Explanations that involve assuming that the producer settles public issues are preferred.

(1) makes it more probable that the explanation is true. One can presumably reduce (2) and (3) to the same principle: it happens to be a fact that it is more likely than not that speakers produce utterances with a maximal degree of coherence and with a maximal degree of relevance. But it would seem that this is not something that is part of the typical resources on which PLAUSIBLE draws: the sort of knowledge known as semantic memory. If it is in fact more likely than not that speakers express maximally coherent and relevant messages, the reason must be that this is because interpretation is biased towards maximising coherence and relevance: if the speaker would express his message in a signal that allows an interpretation that is more coherent and relevant than the message to be expressed, he would be misunderstood. This bias would be inherited from other kinds of perception and the high probability of speakers intending interpretations to which hearers arrive by perceptual bias is due to an accommodation to the perceptual bias.

(2) has a directly relation to probability: extra material is an extra risk for the whole explanation to be false. That does not give its full effect however. That would be conservatism in perception: one assumes that something is the same as what was given in percpetion just before, unless it cannot be. (3) is based on reasoning about the speaker: one needs to find motives for what the speaker is doing. If she seems to be addressing an issue given in the dialogue, the inference that, at that point, she is dealing with that public issue provides the motive.

The bias to coherence and relevance makes it necessary for the speaker to take special measures to block it where it is unwanted. It follows that the bias is harmless: where it should not apply, the speaker must take care that it does not. If this is so, following the bias increases the probability that the explanation is true.

PLAUSIBLE must be implemented by some set of data learnt from experience. It is important to realise that it is the same resource that would give common sense explanations and could be used as a resource in perception. (2) and (3) however are relatively simple operations on given explanations. (2) constructs identifications and bridges, (3) links between the set of given questions and parts of the interpretation. While the results of (2) and (3) have to pass the test of plausibility, presumably the use of a direct implementation is a more reliable device for increasing probability than to obtain a model by causal learning.

The priorities between the constraints make pragmatic interpretation a proper optimisation problem that is solved by an optimality theoretic system of ranked constraints.

The interpretation system therefore maximises $p(E)$, given the context. *NEW and RELEVANCE both minimise the size of the models of the interpretation, *NEW by refusing to create new objects and RELEVANCE by making the answers to activated *wh*-questions exhaustive.

The priming mechanism as studied in psycholinguistics seems to have exactly the properties that one needs for pragmatic optimisation. PLAUSIBLE is the frequency effect, *NEW is the preference for already used concepts and objects, RELEVANCE the preference for information that has been queried for.

4 Production OT for estimating $p(O|E)$

Boersma and Hayes (2001)'s stochastic optimality theory and the learning algorithm that goes with it almost immediately provide what is needed for this paper. OT on production is the postulate that natural language production can be described by ranking a set of universal constraints. Stochastic OT postulates that the ranking is a question of assigning weights and that actual production is noisy: the closeness of the weights increases the chance that in a production the weaker constraint will be the strongest.

What one predicts other people to do as an interpreter can be described as a more noisy version of one's own production. For a given interpretation therefore, stochastic OT allows a direct estimation of what the probabilities are: run enough productions for the interpretation from one's own constraint system using a large noise factor.

The attractive property for our purposes is that under the OT and stochastic OT assumption learning production from experience is quite possible. Under stochastic OT, the resulting system of weighted constraints also provides a probabilistic model of the frequencies with which productions are produced for an interpretation[5].

The importance of Bayesian interpretation would be precisely the fact that $p(E|O)$ cannot be accurately learnt from experience. The success of OT in the production direction would mean that this is not true for the other direction: it is easy to rank or weight the constraints that define the best utterance for a given semantic input.

A precise estimation of $p(E)$ in a given context is more difficult.

Suppose that one has large set of data, e.g. from a semantically annotated

[5] Goldwater and Johnson (2003) proposes Bayesian learning for OT constraints which gives in effect a harmonic grammar. It may be that an even better approximation to the actual probabilities is possible in this way.

corpus. Within a certain window one can then look at the subsets of literals and look at them as instances of a situation type by abstracting from the objects to which the predicates apply. From this one can compute the probability of the situation types. This would then allow estimations of the probability of the elements of the interpretation and of the interpretation as a whole. These can be corrected by the effects of *NEW and RELEVANCE.

If this approach works, both $p(O|E)$ and $p(E)$ can be estimated and a direct implementation of the search for the most probable interpretation for the utterance can be conducted by dynamic programming. The evidence for priming in interpretation however makes it unlikely that the brain proceeds in this way.

5 Doing Bayesian interpretation inside OT)

The proposal of Zeevat (2009) can be seen as a formalisation of Bayesian interpretation inside OT. The three pragmatic constraints are subordinated to a single constraint FAITH that demands that the utterance is as optimal as possible for the candidate interpretation, something to be decided by an OT production system. The demands that should be imposed on proper interpretations (it should be fully resolved and contain the intention of the speaker) are part of GEN.

FAITH > PLAUSIBLE > *NEW > RELEVANCE

The first loop of the algorithm that was described in section refbrain can also be formalised in an OT system. Now it is however necessary to spell out the concept of an interpretation.

COMPLETE & COHERENT > WORDS > PLAUSIBLE > *NEW > RELEVANCE

COMPLETE demands that the set of concepts is resolved: there are links for each of unsaturated slots in the concepts into other material given by the other concepts and the context. COHERENT demands for an attribution of an intention to the speaker and further demands that its predication is the "top"of the set: it connects to everything else (or parts of it) by the transitive closure of the linking relation. WORDS is the requirement that every word (morpheme, multiword expression) contributes a concept with which it is associated (a constraint that allows exceptions, e.g for expletives).

WORDS below COMPLETE and COHERENT allows the input to be defective in the sense that some words can be unrecognised or can be assumed to be misrecognised. It makes the winner the pragmatically best interpretation that supplies a correction to the input.

The system

FAITH > COMPLETE/COHERENT > WORDS > PLAUSIBLE > *NEW > RELEVANT

defines a robust parser. If FAITH is fully met, i.e. the utterance as perceived is an optimal realisation of the interpretation, COMPLETE/COHERENT and WORDS are also fully met. But if FAITH can only be met as well as possible because of defective pronunciation, noise, misrecognition, speaker error and the like, the lower constraints enforce a strategy of going for a full interpretation nonetheless in which the gaps and mistakes are corrected and a maximum number of lexemes are taken into account.

Boersma (in a recent project proposal for a collaboration between ACLC and ILLC at the University of Amsterdam) has presented a programme of doing all of NL processing inside OT. That raises the question whether PLAUSIBLE can be turned into OT. This can be done, very much in the spirit of Boersma (2001). One takes the empirically captured recurrent situation types as constraints ordered by stochastic OT on the basis of their frequency in the corpus of situations. They assign an error to any new context in which the situation type is not instantiated[6]. This model of PLAUSIBLE can be extended to full pragmatics by putting *NEW and RELEVANCE below it. Plausibility can be OT, but it is not linguistics and an OT system of this kind should have applications outside natural language.

6 Improving production

Above it was noted that simulation improving understanding is matched by a similar improvement of production, made possible by the upwards and downwards arrows between levels of representation. If one thinks that it is plausible that simulated production is recruited by the understanding system to improve understanding, it becomes very likely that simulated understanding has been recruited to improve the probability that the hearer will understand the production correctly. The principle is the same, if the lower representation built in formulation does not prime the higher representation it is inhibited.

Simulation checks priming, and priming can check the transformation between representations by testing whether the hearer will get it right, by simulating interpretation. This can lead to early detection of problems and hidden and overt self-corrections and will after the production of the utterance merge with the direct monitoring of the hearer reaction. But there are some descriptive problems that suggest that part of it happens as part of the production process itself and may well be automatic parts of the process. These are cases where production OT by itself fails to give a proper explanation.

The simplest case is optional marking. Lascarides and Asher (1993) give the well-known example (4)where everybody seems to interpret Bill's pushing as the cause of John's falling.

(4) John fell. Bill pushed him.

This example is matched by examples like (5) which receive a prefered interpretation where Mary's smiling is her reaction to John's falling.

(5) John fell. Mary smiled at him.

[6] This can be restricted to situation types that are primed by the new utterance only to keep the number of errors reasonable.

But while a smile is not a very likely cause of a fall, it can be: just assume that John is a beginning ice skater and rather shy. If this is the intended interpretation the speaker should overtly mark it with a causal marker, an optional marker given (4) .

(6) John fell. Because Mary smiled at him.

The marker is obligatory in the case described: when the default interpretation is not causal. Speakers make the decision effortlessly and unconsciously. The question is how the production routines can figure out what is the default interpretation. The problem can be solved by full online bidrectional competitions as assumed by e.g. Blutner (2000), but that position has serious problems (Hale and Reiss, 1998; Beaver and Lee, 2003). Much easier is the assumption that simulated understanding checks whether the causality feature has been realised in a recoverable way. Notice that this is a large problem. Most discourse relations are only optionally marked, additive and adversative particles typically are optional markers of what they express, case marking and tense and aspect marking can be optional in languages and the structure of the problem seems to be exactly the same. The features expressed are important for communicative success and simulated understanding is checking whether the hearer will get them right, so that the marker can be inserted.

The second case is word order freezing[7]. The standard case is Jacobson's (7) (Jacobson, 1984) that can only be interpreted in the way indicated and not as "the daughter loves the mother". Russian generally allows Object-Verb-Subject-sentences, but not in this case where both *doc'* and *mat'* exhibit case syncretism: the nominative and the accusative coincide. The phenomenon has by now been attested in a wide variety of languages.

(7) Mat' ljubit doc'.
The mother loves the daughter

It can be described as a competition between two constraints: canonical subject object order and a constraint that creates sentences with the reverse order, e.g. a principle that fronts contrastive topic. Neither principle can outrank each other, but in the syncretic cases, canonical word order is the clear winner.

In Dutch the following triplet is possible.

(8) Wie ziet Maria?
(ambiguous)Who sees Maria? Who does Maria see?
Maria ziet Jan
(one reading only) Maria sees Jan.
Hem ziet Maria
(one reading only by case marking) Maria sees him.

In the first case, word order is hijacked by the question marking system and it remains unmarked who is the subject and the object. In the second case,

[7] Zeevat (2006) gives a fuller presentation.

the subject and object are marked by word order and (3) illustrates that word order does not always do so, since a case marked object can be fronted to mark contrastive topic.

Pure production does not do the trick. Full online bidirectionality with our two constraints makes the prediction that *Wie ziet Maria* is unambiguous and predicts that the object in the frozen *Maria ziet Jan* cannot be the contrastive topic[8]

A solution by automatised simulated understanding works adequately. Simulated understanding can see that the functional roles are misrecognised. There is no alternative for *Wie ziet Maria?*, the role assignment is correct in the third example and in the second example the competing alternative word order is discarded. The influence of the checking on the output is zero if production has no better alternatives on offer, which suggests that one is dealing with standard production optimisation in freezing. There is also a clear preference for checking the assignment of functional roles over the assignment of contrastive topic, which suggests that the features checked are ranked by their relative importance.

A beautiful example of the same kind has been provided by A.Teodorescu (2006). The word order "Italian tall student" is marked: it should be size before nationality. But in (9)this is overridden presumably because it means something else than "tallest Italian student". Again monitoring should be able to prevent the word order constraint on adjectives to put *tallest* before *Italian*[9]

[8] Bouma (2008) contains a interesting attempt to solve these two problems using bidirectional stratified OT. The idea is that there can be other constraints in the same stratum with the two constraints assumed here. E.g. a definite subject can be better than the indefinite subject *wie*. The best of this family would be *SUBJ/WIE which deals with the ambiguity ruled out by Bouma of *wie trof een steen. Who hit/was hit by a stone.* This makes the prediction in the interpretation direction that sometimes the first element is the subject and sometimes the other NP. The problem is that a constraint of this kind seems hard to learn: *wie* is as often a subject as other NPs. Bouma's approach is similar to Boersma's solution Boersma (2001) to the Rat/Rad problem: use other constraints to occasionally restrict the noxious effect of the production constraint in the interpretation direction. The problem of allowing CT-readings for the object in frozen sentences is also solved in the same way: an extra constraint allows the context to designate the object as the contrastive topic without syntactic marking. It is possible to patch up bidirectional OT in this way to deal with any counterexample that comes up. But is it convincing? Using the reverse competition with the same constraints was meant to provide an account of semantics within OT with no extra effort. The problems with BIOT just show that at least one constraint per problem should be added, i.e. that something was missing in the proposed semantic account. Also, Boersma and Bouma want to convince us that ambiguity does not arise because two messages are mapped onto the same form, but because of extra constraints. This conflicts with a long tradition and with intuition. Boersma's solution brings in semantic constraints to override the reverse effects of a production constraint and that is valuable: semantics is important and the proposal interesting, but there is something wrong with the reason for their introduction. Similarly, the use of markedness constraints on indefinite and inanimate objects is valuable (they have typological significance and a descriptive role to play in Dutch), but to introduce them for saving a "lost" ambiguity is the wrong reason. One first loses the ambiguity by insisting on reverse optimisation and then invents constraints to recover them. Better not lose the ambiguities in the first place.

[9] This is not necessarily a metaphor. In the proposal of Zeevat (2008) production constraints are procedures that enrich underspecified structures if this can be done. The natural interpretation of automatic self-monitoring in such a system would be as riders on the constraints

.

(9) I have an Italian tallest student.

The last two examples are phonological. Boersma (2007) gives a convincing argument that the phonological effects of the silent h in French cannot be captured by production OT. He also formulates and rejects a simple solution in terms of "hearer-oriented max-constraints", exactly what I am proposing for optional marking and freezing. Within the framework of this paper, the existence of cases where production OT should be improved by simulated understanding looking at the understanding of specific features makes good sense. That solution is simpler than the proposal Boersma ends up with.

The final case is articulatory. If one assumes that the production process bottoms out in a sequence of instructions to the articulatory organs that ascribes the goal to the speaker of just carrying out that plan. Instead, empirical findings seem to indicate (Perrier, 2005) that the goal of the speaker has both acoustic and articulatory properties. This again suggests automatised simulated understanding militating against variant realisations that may lead to confusion.

Pure mono-directionality does not lead to satisfactory approaches to the five phenomena discussed above and simulated understanding during production would help. Simulated understanding in production is very plausible once one has assumed simulated production in understanding.

If Bayesian understanding of utterances is implemented in the way described in section 2 it must be an exaption of some other Bayesian understanding process. That must have come into being because of the improvement it brings to understanding: it would not possible in that area to have a reliable direct estimate of $p(E|O)$ and quality goes up by going Bayesian. If matters are like that, simulated understanding in production has become possible by the same architecture. The pragmatic biases lead to frequent misunderstandings precisely because they reflect the most probable message that is consistent with the signal: whenever the user does not want to express the most probable message. The success rates of communication go up dramatically by simulated understanding, if there are expressive devices available that can deal with the problem. The gain in communicative success by simulated production in understanding and by simulated understanding in production is quite comparable and it is hard to imagine that evolution would have selected one without the other. In the model they work by the same principle: a failure to get back the higher representation from the induced lower representation inhibits the process and allows other lower representations to be produced.

One further point to be made here is that other lower representations must be available. Language evolution seems to have created the whole functional inventory with precisely this purpose in mind: to provide marking devices to use when simulated understanding inhibits the unmarked forms. They sit in special areas: tense, aspect, case, topic, number, definiteness, additivity, adversativity, confirmation and others. In all of these cases, PLAUSIBLE and *NEW

that can be harmful to interpretation: do this, unless so and so.

will select the most frequent or oldest possibility for the unmarked form: the present, perfective, subject for animate, object for inanimate, keeping the topic, singular, definite, non-additive, non-adversative, non-confirmation and so on. The functional inventory builds the marked forms precisely in the cases where the pragmatics leads to the non-intended readings.

7 Consequences for Acquisition

With lots of provisos the model also leads to interesting predictions about acquisition. In the beginning, one must assume that there is some understanding based on $p(E)$ and observed language use. These provide the necessary learning data for learning production. Simulated production is not an option at that point, because there is no production. It follows that understanding must still be defective. Early production can therefore not build on simulated understanding because the understanding is not sufficiently accurate yet. But early production can be used to boost understanding after which understanding will in turn boost production. Automatisation can only occur when things are relatively stable.

It seems most likely that simulation of understanding in production and of production in understanding do not happen as one single event but rather per area in the language. In fact, the model predicts that the inhibitory effect when the circle does not close will exist from the very start but should not be very strong. It also will not have much effect due to the difficulty of finding alternative productions and interpretations.

But one can still assume an order in which things happen:

understanding
learning data from observation
production
simulated production in understanding
simulated understanding in production
automatisation

This order can be used to explain the effect of delayed principle B studied by Hendriks and Spenader (2006) and Mattausch and Gülzow (2007) in full online bidirectional OT. The phenomenon is that young children go through a phase where their production of reflexives and pronouns is correct, but in which pronouns still receive reflexive interpretations. (10) can be interpreted as "John hits himself".

(10) John hits him.

The treatment of reflexives in Hendriks and Spenader (2006) is formally correct and even the explanation seems on the right track. But the choice of the constraint set seems arbitrary and there are many alternatives that would not work. E.g. adding a constraint *PRONOUN/DISJOINT would spoil the explanation completely. And that constraint seems as well motivated as the constraint used: *REFLEXIVE/NONLOCAL. Mattausch also brings up the issue that it is hard to say which of reflexives and pronouns is the more marked expression (and reaches the opposite conclusion). It would seem to me that a solution should not depend too much on particular choices of constraints, especially not on decisions about which constraints are not present in learning the ordering.

The treatment of Mattausch and Gülzow (2007) contains a statement of speaker optimality that is hard to stomach: apparently one must monitor one's own production by doing reverse optimisation while the hearer just inverts productive optimisation for interpretation: it is monitoring with the wrong notion of interpretation. This just does not make sense and invalidates the otherwise very ingenious solution[10] The problems with the treatments show that it is not easy to explain the effect in BIOT.

The suggestion contributed here is that the effect should be attributed to the sequence above. In the beginning pronouns are just associated with the concept "referent is highly activated". Subjects in the same clause are also highly activated. In production at the same time a preference for reflexives has been established if the antecedent is the subject in the same clause (just use any constraints that would characterise the preference, nothing depends on it). If simulated production sets in too early, it harms understanding by eliminating correct interpretations: production is not correct yet. So the strength of the inhibitory effect of not getting the lower representation back from the higher representation should be low at this point. When production is more reliable, simulation is beneficial and the strength of the inhibition can be increased. At this point, the principle B effect is achieved. This explanation is more robust and especially simpler than the two BIOT solutions.

The considerations in this section make a number of other predictions. One is that monitoring can start working only after simulation in understanding. Another is that the sophisticated incremental syntax-driven parsing algorithms that come out of the fMRI studies of human processing (Bornkessel and Schlesewsky, 2006) can only be built very late: they are automatisations of a full system of understanding incorporating simulated production and the effects of monitoring. These predictions deserve further study.

8 Conclusions

1. If Bayesian interpretation has formed because it is problematic to work directly with $p(E|O)$, this needs an explanation, especially in the light of the relatively good results of probabilistic parsing. The human learner does not have access to the same amount of data perhaps, or the human processor to dynamic programming or to methods of computing the integrated probabilities of complex new utterances. Perhaps also the developers have come up with bright ideas that go beyond what can emerge in the brain. It is also possible that

[10] From the Bayesian perspective, OT evolution simulation should use production OT for production and interpretation and incorporate the frequency data in a model of priming, to give models of speaking and listening.

there is a threshold that probabilistic methods working with $p(E,O)$ estimation cannot cross.

2. This paper takes the line of Hale and Reiss (1998) seriously that OT defines the relation between forms and meanings in production optimisation. It should be taken seriously, because it is the correctness criterion for descriptive work in production OT. That line has however three problems. First of all, one can translate an OT problem and its constraints into a processing model as in e.g. Frank and Satta (1998); Karttunen (1998) and Zeevat (2008) or take it as a high level specification formalism to be cashed out by processing using low level compiled grammars as in Kuhn (2003).

That means that the line of Hale and Reiss (1998) needs a processing model in the other direction where that is not directly provided. Notice that Karttunen's proposal in fact both computes outputs from inputs and inputs from outputs, where the return journey is not using inverse optimisation, i.e. it is an approximation to Hale and Reiss's proposal. But this is not true for Zeevat (2008) for syntax and may well be the case for other approaches to constructive optimisation that take OT seriously. A general processing model for the road from output to input is therefore missing. A related problem is that learning needs some mechanism to obtain learning data.

The second problem is semantic blocking (production OT can do productive blocking). This happens e.g. in the example (11)I used before.

(11) Katja and Henk were surprised that the editors rejected each other's papers.
Katja and Henk were surprised that the journal rejected each other's paper.

The third and perhaps most interesting problem is that it seems that pure mono-directionality faces empirical problems: not everything can be described by production OT systems as discussed in section 6. This paper provides a principled solution to all three problems. The empirical problems can be addressed by assuming integrated simulated understanding in production. Interpretation can start with priming and pragmatics and finish by simulated production. This solves the first and the second problem. Simulated interpretation in production and simulated production in interpretation both add bidirectionality to the model, but it is not the bidirectionality of Smolensky (1996) or Blutner (2000) in which interpretation is equated with the reverse optimisation where the productive constraint system chooses the best interpretation given the form as input.

3. The problem that Hale and Reiss (1998) identified with the rat/rad-problem is not a problem of bidirectionality, which from the perspective of this paper is a profound insight in human languages and how they are used, learnt and develop but a problem with reverse optimisation. From the perspective of this paper, a production constraint system is the best candidate on the market for a theoretical model of $p(O|E)$ if the ranking or weighting is a result of the learning data: how often do Es lead to which Os. Why would a model of $p(O|E)$ be

able by itself to do interpretation? The chance that it is any good as a model of $p(E|O)$ is negligeable.

The integrated constraint system would do better as a model of $p(E|O)$ if it tried to fit the interpretation data, but would it then still be good as a model of $p(O|E)$? Quite the same holds for bidirectional learning: one expects an empirical fit that is not as good as it can be[11].

Bidirectionality is a central concept for accounts of speaking, hearing, language learning and language evolution, but reverse optimisation of production OT does not help at all from the Bayesian perspective.

4. How does the model presented here compare with other traditions in formal grammar? And their probabilistic versions? In the latter case one would want to start with demanding separate versions for parsing and generation.

Some of the assumptions are in conflict: e.g. the idea that production is learnt and then plays the decisive role in interpretation is in conflict with formalisms like LFG and HPSG (or many variants of CG) which have been designed with parsing in mind and have turned out to be applicable to production only with difficulty. (One can say that proposals for OT-LFG ((Bresnan, 2000) and many others) turn this around and the idea of using OT-LFG to abstractly specify classical LFG grammars is very interesting). Perhaps one should develop an OT-HPSG.

Production OT and OT-LFG are much like early transformational grammar in constraining the mapping from conceptual structure to surface forms. (It would seem that systems of constraints are better at the job and considerably more explanatory.)

The interpretation of grammar as just a constraint on the relation between forms and meanings runs foul of the ambiguity problem. Words and constructions are ambiguous which leads to an exponential number of readings over some base > 2. Variation in production may be smaller, but one would expect it to be above 1. This makes the probability of successful communication if only grammar is used $r^k + s^k$ where r is the base for interpretation and s the one for production and k the length of the message. Non-probabilistic grammar merely raises the problem how signals of any length (e.g. this paper) can bring about communication and does almost nothing in solving the problem. In terms of the terminology of Liberman and Mattingly (1985), the parity problem is ignored.

Probabilistic versions of classical grammar do quite a lot better, in fact better than they should if Bayesian interpretation has come about by evolutionary pressure. Why should people emulate Bayesian interpretation if making classical grammars probabilistic is already the answer to the problem[12]? The same

[11] The best result in this respect is Boersma's approach to the rat/rad-problem Boersma (2001). The constraint system is extended by a set of interpretation constraints that do not influence production. Bidirectional learning will then not disturb the ranking of the production constraints too much and the interpretation constraints patch up the cases where just using the production constraints would lead to the wrong results in the interpretation direction. It is best considered a variant of the model proposed in this paper, where production and interpretation are learnt separately.

[12] In these cases, the underlying grammar can be taken to be a rough model of $p(O|E)$,

question applies to purely probabilistic approaches. Why did people go Bayesian if it was not necessary or at least helpful?

5. The road to making the considerations in this paper into working technology is a long but feasible one. One needs good production OT, preferably for one whole language. Notice that the descriptive problem is considerably simplified because recursion has been relegated to the input. Proposals for probabilistic grammar and word sense disambiguation are good enough to simulate the priming process. The same proposals can be put to the task of the estimation of semantic plausibility by analysing text. In the last case, it becomes necessary to make contact with NL semantics and discourse to obtain an adequate logical model of the concepts, linking and linking preferences.

6. Panini, the founder of the field of linguistics, reduced syntax to morphology, Gil (2005) claims that there languages without syntax and scepticism about the importance of syntax is more common than one would think. One of the conclusions of this paper is that it is important for interpretation in the sense that $p(O|E)$ should have only a few peaks for a given E. This conclusion extends to language evolution: regimentation will be selected by evolution because it is better for understanding to have only a few peaks. It follows that it should also hold for Gil's Riouw Islands Malay, it just does not show up in anything that linguists recognise as syntax. This should be testable by doing interpretation tests on both the Malay in question and other languages.

Bayesian interpretation includes $p(E)$-estimation and thereby denies directly that an account of a language can be autonomous. $p(E)$-estimation has to do with the kind of things people say to each other and the kind of events that happen in the world and the frequency with which they happen. By simulated understanding in production, $p(E)$-estimation influences the description of what one should say in a given situation. And this has its reflex on the functional inventory of languages: evolution has constructed it so that we can use it.

7. As Grice (1975) has it, pragmatics is cooperation, a line that has been taken further in e.g. relevance theory or the bidirectional notions of Horn (1984), Levinson (2000) and Blutner (2000). The break-through in the analysis of presupposition is Heim (1983) (rediscovered by another road by (Van der Sandt, 1992)) that establishes two preferences: for resolution over accommodation and for global accommodation over local accommodation. To the persistent regret of Heim, she was not able to reduce these preferences to Gricean pragmatics or to one of its successors. Blutner was able to reformulate the preferences as OT constraints, making the first preference come out of a semantic economy principle ("minimise the number of new discourse referents") and the second out of a principle that maximises the informational strength of the interpretation. In my reformulation, these have become *NEW and RELEVANCE and Zeevat (2009) shows that they can be taken as an account of general pragmatics that can account for implicatures, pronoun resolution, presupposition and rhetorical

which explains why in some respects these models are better than pure probabilistic ones. Correctness is however harmed because the underlying structure to the problem of learning interpretation, i.e. the objects that are weighted in learning are dictated by the formal structure of the grammar and do not guarantee the best empirical fit.

structure. The two steps forward with respect to Gricean pragmatics are that presupposition now falls out of general pragmatics and that rhetorical structure can be incorporated. The account has a natural interpretation as the optimisation of the explanation of communicative acts by other agents. PLAUSIBLE makes the explanation as true as possible, *NEW is Ockhams razor and RELEVANCE brings in the motives of the other agent and cooperation. This view of pragmatics as optimising explanations has been pioneered by Jerry Hobbs and by reducing also the motivational aspects and parsimony to weighted abduction, he can also be held to be the inventor of the idea that pragmatics is about the maximisation of $p(E)$. The final step is to say that pragmatics is priming. The signal primes the hearer for words and morphemes and concepts and their links and the objects that bind them and for the motives of the speaker. The current brain activation of the objects that can be primed for decides what comes up first and depends on relevance, recent use, general frequency and plausibility. The interpreting brain is a pragmatic machine and not just when it is doing language, but also when it perceives natural phenomena and the actions of others. It needs to get the best explanation of its environment and will recruit simulation whenever it can to make the explanations even better.

References

A.Teodorescu (2006). Adjective ordering restrictions revisited. In *Proceedings of WCCFL*.

Beaver, D. and Lee, H. (2003). Input-output mismatches in OT. In Blutner, R. and Zeevat, H., editors, *Pragmatics and Optimality Theory*. Palgrave.

Blutner, R. (2000). Some aspects of optimality in natural language interpretation. *Journal of Semantics*, 17:189–216.

Boersma, P. (2001). Phonology-semantics interaction in ot, and its acquisition. In Kirchner, R., Wikeley, W., and Pater, J., editors, *Papers in Experimental and Theoretical Linguistics*, volume 6. University of Alberta, Edmonton.

Boersma, P. (2007). Some listener-oriented accounts of h-aspir in french. *Lingua*, 117.

Boersma, P. and Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32.

Bornkessel, I. and Schlesewsky, M. (2006). The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological Review*, 113:787–821.

Bouma, G. (2008). *Starting a sentence in Dutch: A corpus study of subjectand object-fronting*. PhD thesis, University of Groningen.

Bresnan, J. (2000). Optimal syntax. In Dekkers, J., van der Leeuw, F., and van de Weijer, J., editors, *Optimality Theory: Phonology, Syntax and Acquisition*, pages 334–385. Oxford University Press.

Clark, H. (1996). *Using language.* CUP, Cambridge.

Frank, R. and Satta, G. (1998). Optimality theory and the generative complexity of constraint violability. *Computational Linguistics*, 24(1):307–315.

Galantucci, B., A., F. C., and Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13:361–377.

Gil, D. (2005). Word order without syntactic categories: How riau indonesian does it? In Carnie, A., Harley, H., and Dooley, S. A., editors, *Verb First*, pages 243–263.

Goldwater, S. and Johnson, M. (2003). Learning ot constraint rankings using a maximal entropy model. In Spenader, J., Eriksson, A., and Dahl, O., editors, *Proceedings of the Stockholm workshop on Variation within Optimality Theory*, pages 111–120. Stockholm University.

Grice, H. (1957). Meaning. *Philosophical Review*, 67:377–388.

Grice, P. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Syntax and Semantics 3: Speech Acts*, pages 41–58. Academic Press, New York.

Hale, M. and Reiss, C. (1998). Formal and empirical arguments concerning phonological acquisition. *Linguistic Inquiry*, 29:656–683.

Hamm, F. and van Lambalgen, M. (2005). *The Proper Treatment of Events.* Blackwell.

Heim, I. (1983). On the projection problem for presuppositions. In Barlow, M., Flickinger, D., and Westcoat, M., editors, *Second Annual West Coast Conference on Formal Linguistics*, pages 114–126. Stanford University.

Hendriks, P. and Spenader, J. (2005/2006). When production precedes comprehension: An optimization approach to the acquisition of pronouns. *Language Acquisition: A Journal of Developmental Linguistics*, 13:319–348.

Hobbs, J., Stickel, M., Appelt, D., and Martin, P. (1990). Interpretation as abduction. Technical Report 499, SRI International, Menlo Park, California.

Horn, L. (1984). Towards a new taxonomy for pragmatic inference: Q-based and R-based implicatures. In Schiffrin, D., editor, *Meaning, Form, and Use in Context*, pages 11–42. Georgetown University Press, Washington.

Jacobson, R. (1958/1984). Morphological observations on Slavic declension (the structure of Russian case forms). In Waugh, L. R. and Halle, M., editors, *Roman Jakobson. Russian and Slavic grammar: Studies 1931-1981.*, pages 105–133. Mouton de Gruyter, Berlin.

Kamp, H. (1981). A theory of truth and semantic representation. In Groenendijk, J., Janssen, T., and Stokhof, M., editors, *Formal Methods in the Study of Language, Part 1*, volume 135, pages 277–322. Mathematical Centre

Tracts, Amsterdam. Reprinted in Jeroen Groenendijk, Theo Janssen and Martin Stokhof (eds), 1984, *Truth, Interpretation, and Information; Selected Papers from the Third Amsterdam Colloquium*, Foris, Dordrecht, pp. 1–41.

Karttunen, L. (1998). The proper treatment of optimality in computational phonology. In Oflazer, K., editor, *Finite State Methods in Natural Language Processing*, pages 1–12, Bilkent University.

Kilner, J. M., Friston, K. J., and Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive processing*, 8(3):159–166.

Kuhn, J. (2003). *Optimality-Theoretic Syntax: A Declarative Approach.* CSLI Publications, Stanford.

Lascarides, A. and Asher, N. (1993). Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16:437–493.

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

Liberman, A. and Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition*, 21:1–36.

Mattausch, J. and Gülzow, I. (2007). A note on acquisition in frequency-based accounts of binding phenomena.

Perrier, P. (2005). Control and representations in speech production. In *ZAS Papers in Linguistics 40*, pages 109–132. ZAS, Berlin.

Schulz, K. (2007). *Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals.* PhD thesis, University of Amsterdam.

Smolensky, P. (1996). On the comprehension/production dilemma in child language. *Linguistic Inquiry*, 27:720–731.

Sperber, D. and Wilson, D. (1986/95). *Relevance: Communication and Cognition.* Basil Blackwell, Oxford.

Van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *Journal of Semantics*, 9:333–377.

Van Rooy, R. (2003). Relevance and bidirectional OT. In Blutner, R. and Zeevat, H., editors, *Pragmatics and Optimality Theory*, pages 173–210. Palgrave.

Zeevat, H. (2001). The asymmetry of optimality theoretic syntax and semantics. *Journal of Semantics*, 17(3):243–262.

Zeevat, H. (2006). Freezing and marking. *Linguistics*, 44-5:1097–1111.

Zeevat, H. (2007). Optimal interpretation for rhetorical relations. ms, University of Amsterdam.

Zeevat, H. (2008). Constructive optimality theoretic syntax. In Villadsen, J. and Christiansen, H., editors, *Constraints and Language Processing*, pages 76–88, ESSLLI Hamburg University.

Zeevat, H. (2009). Optimal interpretation as an alternative to gricean pragmatics. In *Structuring information in discourse: the explicit /implicit dimension*, Oslo Studies in Language. OSLA, Oslo.