Applying Bayesian Interpretation to Learning

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

1 Introduction

The literal meaning of Bayesian interpretation would be an interpretation method based on two models: one of generation probabilities $p(u|i)$ (the probability of an utterance given a message, meaning or interpretation) and the probability $p(i)$ of the message, meaning or interpretation i itself. (Both models must be context dependent, i.e. $p(i)$ should really be $p(i|c)$ and $p(u|i)$ $p(u|i\&c)$, but the dependence on c will not be in the notation in the following.).

By Bayes's theorem the most probable meaning of an utterance u, $max_i p(i|u)$ equals the interpretation for which the product of $p(i)$ and $p(u|i)$ is maximal, i.e. $max_i p(i|u) = max_i p(i)p(u|i)$.

It will be argued that the human interpretation mechanism is an emulation of Bayesian interpretation based on the capacity to simulate utterance production, the ability to rank hypotheses by plausibility and on direct cue-based association based on the utterance. Cue-based association ("the direct method") by itself cannot reach the accuracy of human beings, as argued in section 2. Section 3 explains the model and gives some arguments for it.

The point of this paper is to apply the model to first language acquisition in section 4. Bayesian interpretation (and the corresponding production model) or rather its human emulation turns out to impose a natural order on the different stages of acquisition and is able to deal with production-comprehension asymmetries in acquisition.

2 The ambiguity problem and the direct method

The ambiguity problem forms the motivation for stochastic computational linguistics. Rule based grammar, a treatment of the lexicon in which all readings of a word are treated equally and pragmatics without preferences lead to a very large set of interpretations for a given utterance. Any single reading has only a very small chance of being the right one. Human interpreters are rarely aware of other readings however and are overwhelmingly right in their interpretations. A proper treatment of the interpretation of NL utterances must therefore aim to explain not just what interpretations are possible but try to predict what interpretation a human interpreter would obtain for the utterance. This interpretation can be equated both with the interpretation the speaker intended and with the most probable interpretation in the given context. If the speaker intended a interpretation that is not the most probable in the context, she would have noticed it and would have produced a different utterance instead. The speaker is also an interpreter and can predict what the interpreter of the utterance would do.

The ambiguity problem is rarely taken in its full generality. It must be divided in a number of subproblems.

1. Sound signal to strings of words (speech recognition)
2. Strings of words to syntactic structure (parsing)
3. Word to concept (lexical disambiguation)
4. Syntactic structure to logical form (especially *scope* of operators)
5. Context integration (*resolution* of pronouns, presupposition triggers, demonstratives, particles, nouns, names, tense, discourse relations, etc.)

For each of these, there exist computational approaches to disambiguation.

Speech recognition is —like other signal processing— wrought with uncertainty, even in the presence of high quality physical measurement and abundant training data. Due to the nature of the data, the standard method in the area is Bayesian: a language model maximises the likelyhood of the message and an articulation model computes how likely it is to lead to the message. A considerable degree of uncertainty however remains.

The large number of syntactic ambiguities generated by large coverage grammars has been effective in generating a reorientation in *parsing* from logical methods to stochastic methods in which finding the most probable parse (given a corpus) is the aim. These methods differ from classical parsing in being full-coverage and in coming up with a most preferred parse.

Lexical ambiguities are an important source of ambiguity. The standard wisdom is that these should be eliminated using the context, and there are implemented systems.

It is nowadays generally assumed that there are various ways of turning a syntactic structure into a *logical form.* It may be possible to use stochastic methods here as well.

One aspect of *context integration* has been well studied: pronoun resolution. Various approaches have been developed, including stochastic ones. Pronoun resolution is however only a small part of the real problem of context integration, which has not been explored in its entirety from a stochastic perpective.

In all these cases, probabilities have to be extrapolated and cannot be determined by just counting: there are infinitely many speech signals, sentences and contexts and even if finiteness could be assumed (e.g. by assuming a maximum length), the number of available data will be insufficient for directly determining the probabilities due to data sparseness. The methods first make the number of events sufficiently small for allowing counting (data sparseness does not go away but becomes managable) and formulas compute the probabilities for the larger events. While theoretically, this can be just the right thing, in practice it does not seem that it is: there are considerable error rates that get eaten away ever more slowly in ongoing research, in the well-researched areas like speech recognition and parsing.

The problem is not that the individual modules lack quality or applications. Where they exist, they are the best methods there are. There is however a problem when they should be combined into a single method for disambiguation.

Assume that each module has a success rate of 0.9 (a realistic figure for the existing systems). Then that gives a success rate of 0.58 only for the composition of all the systems. And it is not clear how that figure can be improved, apart from improvements in the components.

If humans would interpret utterances by this direct method, they would run into the same problems. There is no reason to assume that they have miraculous capacities of stochastic estimation or more data to work with: it is quite the reverse. The stochastic methods employ vast data sets and have very sophisticated mathematics. If humans were using the direct method, they would be well below the 58% success rate and they very clearly do better.

3 The Model

It is much more likely to assume that while humans can use the composition of the direct estimates as a heuristic tool in achieving probable interpretations, they are in fact boosting the probability of their interpretations by using the Bayesian strategy, i.e. by finding maxima for the product of the probability of the interpretation in the utterance situation and the probability of the utterance given the interpretation in the utterance situation. They are good at estimating the probabilities of natural events (a crucial skill in perception and deciding on action) which makes it possible to compare the content of messages for their probability. And crucially, they can reliably estimate the probability of the utterance given the interpretation by simulated language production.

Given the problems noted above, this is the only way towards a more accurate estimation of the most probable interpretation.

The direct method can be used for selecting the most promising candidates and can also be used as the baseline for the Bayesian method. In a case where the direct method has selected the correct interpretation, the interpretation will also have a maximal product of the two probabilities, i.e. likely competitors must be less probable or the utterance must be worse as a formulation of them than the winner. It is therefore highly improbable that the correct interpretation will be deselected by the integrated algorithm.

The proposed algorithm makes two assumptions. The first is that the estimation of the probability of the utterance given the interpretation can be reduced to a classification in high, lower, low and very low probability, corresponding with correct, correct but unusual, incorrect but recoverable and incorrect. Further, that the interpretations can be ordered by their a priori probability. Both assumptions are weaker than the assumption that the brain can do proper numbers for both probabilities.

The algorithm is given as follows.

1. Obtain the $k(u)$ best results of the direct method.
2. Consider only the best results in terms of the generation probability classification (only the high probable ones, if there are any, else only the lower probable ones, if there are any, else only the low probable ones, if there are any, otherwise all the results).
3. Pick the most probable interpretation from the resulting set.

Assume that the direct method is fully available, i.e. there are procedures $DM(u)$ and $DM_{k(u)}(u)$ such that the likelyhood that $DM(u) = max_i(p(u|i)p(i))$ is reasonably high (below 0.58, but still substantial) and the likelyhood that $max_i(p(u|i)p(i)) \in DM_{k(u)}(u)$ is very high. This would seem to be a matter of selecting $k(u)$ appropriately high.

Now further assume that $DM_{k(u)}(u)$ can be accurately ordered by $p(i)$ and classified by $p(u|i)$.

(1) $BM(u) = max_{i\in\{i\in DM_k(u)}p(u|i)$ is high$\}p(i)$ if defined
otherwise $max_{i\in\{i\in DM_k(u)} : p(u|i)$ is lower$\}p(i)$ if defined
otherwise $max_{i\in\{i\in DM_k(u)} : p(u|i)$ is low$\}p(i)$ if defined
otherwise $max_{i\in DM_k(u)} : p(i)$

Now consider a normal speaker, knowing the language and able to monitor her utterance by a simulated interpretation. Having a normal speaker implies that the intended interpretation will score a high value on $p(u|i)$. Monitoring implies that there will not be an equally or more probable alternative interpretation with a large value in $DM_{ku}(u)$. Under these assumptions and conditioned by the likelyhood that $max_i p(u|i)p(i)$ is in $DM_{k(u)}(u)$ and under the assumption that the interpreter does a good job in estimating both $p(i)$ and p(u—i)\$$BM(u)$ is in fact $max_i p(u|i)p(i)$. Assuming 0.95 as the likelyhood of both having a normal self-monitoring speaker and $max_i p(u|i)p(i) \in DM_k(u)$, this will give roughly the same success rate of 0.95 for the whole process, i.e. well above the baseline.

The speaker can however still express herself in an unusual way, make mistakes or be less than fully competent. The reliability of $BM(u)$ drops in those cases. In a conversational setting, these are the cases where the hearer will check the interpretation by producing grounding moves.

One cannot do away with the direct method. In that case one would have to compare all the infinitely many possible interpretations and check whether they are optimi for $p(u|i)$: this is a bad search problem. The search space needs to be trimmed and the direct method is the most rational method for doing that. One cannot start from the most likely interpretations either, since it is definitely possible to say very surprising things which would come up long after the more likely interpretations.

The hypothesis of this paper is that the human brain implements a version of the algorithm above. It implements it as a single associative process that incorporates (a), (b) and (c).

(2) a. beam search for full interpretations on the basis of cues
b. maximising probability of the message in the context
c. simulated generation

(a) recognises cues for activation and treats the activated entities a further cues for new activations. Recognised phonological cues activate words, words activate concepts, concepts activate links to their arguments and components of the given linguistic context and of the background.

An association path to an interpretation is successful if all incompleteness in concepts has disappeared, all the words are used and the combination forms a proper speech act of the speaker.

The results can then be pictured as a combination of a complete dependency graph and a resolution graph which can be interpreted as the representation of the context updated by the utterance and as the intention of the speaker to bring about that update.

(b) biases (a) by frequency, recency and expectation. This is what comes out of priming research: word sense priming is influenced by these factors. It is here generalised to larger units of meaning and to links to components of the context. In 3.3, it is argued that these factors contribute to the a priori probability of the interpretation.

(c) is the same mechanism that transforms a thought into speech in language production. Within the combined process it deactivates candidate interpretations whose preferred expression does not match the perceived signal.

In the following five subsections, arguments for the hypothesis will be presented.

3.1 Bayesian interpretation

As we saw, the mechanism above approximates: Interpret u by i such that $p(i)p(u|i)$ is maximal.

It gives the intended interpretation i iff the speaker has chosen u such that $p(i|u)$ is high, i.e. that the utterance is correct for the interpretation.

It is not infallible: there is no guarantee that beam search will hit the maximum and there can be discrepancies in the estimation of $p(i)$ and $p(u|i)$. But it can be assumed that it is highly accurate under favourable circumstances.

3.2 Evolution

Any kind of perception can be described as the attempt to find the most probable explanation of a signal.

An explanation is finding some state of affairs that would cause the signal. The probability of the explanation is increased with the probability of the state of afairs and with the probability that the state causes the signal.

Perception —but also explanation that is not perception and planning— therefore requires the estimation of the probabilities of states of affairs (in a context). It also requires causal reasoning: how likely is it that the state of affairs causes the signal (the phenomenon to be explained, the goal of the plan).

In the explanation of behaviour of others, the causality can be judged partly by estimating how easily the explanation leads to the observed behaviour by trying to simulate the planning of the behaviour and the behaviour itself.

Since language interpreters can speak, they can simulate the language behaviour they observe.

Interpretation of language is a special case of perception. Humans are very good at it because they combine the estimation of the probability of the interpretation with simulation and perceptual cues.

The evolution of language understanding is then just adapting normal perception to the special case of language, using the already existing model of a priori probability of percepta. The real evolutionary event in the evolution of language (or more appropriately the evolution of dialogue involving human language) is the emergence of language production. The real event in the emergence of language understanding is exapting the emerging production for understanding.

Evolutionary considerations are also relevant for the conditions under which the mechanism is a good approximation to $p(i)p(s|i)$.

Failure can be the consequence of three factors:

(3) a. beam search misses the maximum.
b. there is a distractor that is more probable but has a similar value for $p(u|i)$.
c. discrepancies between the models of the probabilities.

Factor (c) cannot be eliminated and will cause communication errors. But the speaker can see both that the intended interpretation is not activated by beam search and that there are distractors. Speaker self-monitoring is just the postulate that the speaker rejects the utterance in favour of another in both cases.

Syntax makes distractors rare. If there are two interpretations of an utterance that can be distractors of each other, they should preferably not both have a high $p(u|i)$. More abstractly, for a given u, there should be few i such that $p(u|i)$ is high.

Both syntax and speaker self-monitoring will be reinforced in the evolution of language use. Speaker self-monitoring directly pays off in higher communicative success rates. Any formal device that rules out distractors will be promoted by speaker self-monitoring and has a chance of becoming a strict rule.

It is plausible that the interpretation mechanism has worked like this since the start of the evolution of human languages. It follows that it has been a side condition (changing only in the simulation of the emerging production mechanism) under which production has emerged. It should follow that beam search is good enough for the decoding of the productions, it will almost never miss the intended interpretation.

Speaker self-monitoring is likewise enforced by evolution: without monitoring, the communicative success rate is lower. If communication has a survival value (as shown by the emergence of language), it follows that speaker self-monitoring also has survival value. And the resources it requires are just there, it is the interpretation mechanism.

3.3 Linguistics

Production OT is not just a good tool for the description of phonology, morphology and syntax, it can also be turned into a good model of $p(u|i)$. This has been shown by Boersma (see e.g. **?**) and has been adopted in studies of free variation (**?**, **?**). The one criticism of the model is the lack of speaker self-monitoring in production OT (**?**). **?** discusses similar phenomena in phonology, but solves the problem in a different way. Other models can be proposed that provide similar models of $p(u|i)$, e.g. harmonic grammar (**?**, **?**). In such models $p(u|i)$ is accurately estimated by a grammar that is learnt by the learning algorithm that comes with the model.

Second, pragmatics can be reduced to probability maximisation. This is the essence of Hobbs' interpretation by weighted abduction, when the costs assigned to the explanation rules in abduction are interpreted by probability, as **?** proposes.

Within OT, **?** gives a model that proposes three constraints:
PLAUSIBLE > *NEW > RELEVANCE. PLAUSIBLE is about the probability of the message in the context: this should be maximised. It contains two aspects: the prediction of the hearer about what the speaker is going to say at the particular moment in the conversation. The other is the probability of the particular content in the context (*Do pigs fly?*).

The other two principles can be seen as reflecting perceptual biases in normal perception, especially vision: *NEW prefers minimal changes to what is already given and RELEVANCE is the assumption that the interpretation should be taken as settling activated issues (including ones that are activated by the utterance) whenever it can do so. These two principles can be identified with the recency and expectation effects in priming.

Speaker self-monitoring predicts that the speaker will notice these perceptual effects on interpretation and will try to avoid them if they are not intended. It therefore follows from the existence of speaker self-monitoring that the effects increase the probability of the interpretation that they enforce. It is therefore correct to equate this style of OT pragmatics with probability maximation as well.

The argument from linguistics is that models of both components of Bayesian interpretation have come out of linguistic research: the probability of the message in the context and of the probability of the utterance given the interpretation. This in marked contrast to what linguistics has achieved in interpretation by the direct method: there is no comparable model.

It should be noticed that probability maximisation on the content level is not linguistics, but world knowledge. For a proper implementation of probability maximation, it would be necessary to develop a connection between the sophisticated models in knowledge representation in AI (e.g. Bayesian nets for causal reasoning or frames) and the corpus based methods of stochastic computational linguistics.

3.4 Psychological and neurological evidence for hypothesis

Studies about priming were already mentioned. The bias towards recency, frequency and expectation is however not limited to linguistics, but has also been observed in vision and other perception.

There is very considerable evidence for self-monitoring on many levels during production (**?**, **?**, **?**, **?**). For the hypothesis, one merely has to assume that part of self-monitoring happens before articulation and is automatised. It would cover what in OT has been known as expressive constraints or recoverability. **?** proposes to limit automatic self-monitoring to an ordered set of features: the presence of the feature in the input inhibits a production that is also optimal for the same input without the feature. This deselects the production if there is an alternative optimal production which is not also optimal for this competitor.

It is tempting to interpret the mirror neuron systems (**?**) as involved in simulation tasks with the aim of boosting perception. The other purposes that have been proposed for the mirror neuron system seem less convincing: not all species with mirror neurons imitate or learn by imitation. The case where understanding can be equated with simulation seems limited to understanding emotions (these are states of the motivation system and as such cases where simulation exhausts their understanding).

The proposal is to give simulation (the best model of $p(u|i)$) together with probability estimation of the message a role in the improvement of perception and understanding. The direct method gives perception and understanding, but their accuracy (and thereby their scope) is relatively low and boosted significantly if simulation and probability maximation are integrated in the perception and understanding mechanisms in the way sketched above.

The mirror neuron systems seem to do simulation in perception and their evolutionary explanation may well be that simulation —in combination with semantic memory— dramatically increases the quality with which the behaviour of others is perceived. The interpretation of mirror neurons as involved in improving visual perception along Bayesian lines has been adopted by **?**.

Mirror neurons come with the question what functional advantages created them in evolution. This question must have an answer, otherwise they would not exist. Bayesian perception is a simple answer to this question. While other answers may still be forthcoming, the existence of mirror neurons and the Bayesian answer is an additional reason for taking Bayesian accounts of perception seriously.

4 Learning

There are now four components (4) in our model. (a) and (b) are fully integrated into biased beam search and form the implementation of (d). The four components form one module activated in production and perception: biased beam search activates simulation/production and simulation/production activates biased beam search.

(4) a. beam search
b. plausibility maximation
c. simulation/production
d. speaker self-monitoring

Simulation/production in interpretation is an inhibition mechanism: bad match (the interpreter would not say it this way) with the utterance deactivates interpretations.

(5) a+b ← c

Monitoring interpretation (d, using a+b) inhibits production: a bad match (in the sense of not making important input features recoverable) with the input for c deactivates production plans.

(6) c ← a+b

The application to learning is based on two impossibilities: one cannot simulate speaking unless one can speak and one cannot monitor one's speaking unless one can understand.

But this seems too categorical: one can always speak a little bit and understand a little bit. A much better formulation is to assume that the inhibitory links grow stronger with the skills that are acquired to the degree that they improve speaking and understanding. This is normal learning: good experience with the inhibitory links makes them stronger, bad experience makes them weaker. The proposal is that the inhibitory links are always there and grow stronger with the emergence of better speaking and understanding.

The one thing that can be learnt directly are associations between words (or morphemes) and concepts. This is a finite problem[1].

[1] If one takes phoneme learning to be integrated in word learning. Otherwise this would be a similar task of learning associations between articulatory gestures and perceptual cues fromn which these gestures can be recognised.

One should assume that children start with learning such associations passively by hearing the word in the presence of instantiations of the concept. If they have the concept, this builds an association. It is clear however that strong feedback only comes with the attempts to use the word oneself: in speaking one gets what one wants or not.

Word learning happens at a point where perception, explanation, planning, communication and concepts are already in place. Initially, learning is building associations between words and concepts activated when the word is used. Word production offers the possibility of further feedback and what is constructed is a situation where words activate concepts and concepts pronunciations of the word. Once this is in place, simulation in understanding (would I use this word in this situation?) and self-monitoring (will I be understood when I use this word?) can start functioning.

Simulation makes it possible to recover from wrong interpretations (coming from non-linguistic cues which obliterate the association evoked by the word). Self-monitoring and feedback can prompt the mobilisation of alternative means of expression.

Word Combinations

In early word combinations there appears to be no syntactic categories, morphology or meaningful word order, with the possible exception of putting the topic first.

The feedback at this stage is weak: none of these factors is decisive for understanding. The adult input is also not directly helpful: it will consist mostly of words that are not understood.

The emergence of simulation in understanding will provide the learning data for syntax and morphology, of the kind needed for OT learning. The child understands the adult as meaning i which it could express as u' unlike the observed u. This can demote those constraints that make u' a winner.

The emergent self-monitoring will drive the child towards extra complexity i.e. to go for more words.

Multiword utterances are the result with a concomitant boost of simulation, grammar learning and the emergence of morphology and word order. Since morphology and word order do help, simulation at this point leads to an adult level of understanding.

The final step is the development of further self-monitoring leading to strategic "pragmatic" production: NP selection, pragmatic use of particles and connectives.

The picture can be summarised as follows

1. perception, explanation, planning, communication and concepts
2. direct word learning
3. word production
4. word simulation in understanding

5. word combination production
6. simulation for word combinations
7. syntax learning
8. syntactic understanding (with full simulation)
9. self-monitoring on full utterances

Does this picture fit the facts?

Acquisition starts with single words expressing concepts in which the holes have to be filled in from the context of utterance. In the considerations above, the important fact is that direct learning is possible for words. This leads to a table of word-concept associations that can be run in both directions and one word utterances are the result.

Self-monitoring in word production has to emerge for the two word stage: sometimes the context gives the wrong binding. (7) will give better results if the object of *eat* is not the contextually salient porridge.

(7) Eat cookie.

The same mechanism is responsible for longer utterances.

The advent of utterance of two and more words will produce the data for OT learning of syntax and morphology. An understood utterance will be simulated from the understanding, but rather than suppressing it will bring reranking in the syntactic constraint system, eventually leading to a correct production grammar.

This is the point where syntax and morphology start emerging in the output.

There is a lot to be said about this process, but not from the perspective of this paper, but from the perspective of e.g. OT syntax learning. An option is to start with equally ranked constraints that become ranked with respect to each other in the process, so that fewer and fewer productions are going to be permissible.

The emergence of syntax will boost understanding by allowing simulation, so that adult level self-monitoring becomes possible. E.g in wanting to express (8a) (8b) should be suppressed for the same interpretation, even though it is perfectly correct from a syntactic point of view and would even be the preferred formulation for more likely causes of falling.

(8) a. John fell. Because Mary smiled at him.
b. John fell. Mary smiled at him.

A similar example is (9), where (b) is incorrect unless it occurs in a list answer.

(9) a. John fell. Bill fell too.
b. John fell. Bill fell.

According to **?** appropriate use emerges after everything else. Unsurprisingly perhaps, this is also the part of text generation that is still the least understood and makes free text generation as problematic as it still is.

4.1 Application: the reflexive asymmetry

It has been observed (see references in **?**) that young children go through a phase in which reflexives are produced correctly, but where non-reflexive pronouns can be understood reflexively. In this way, these children can understand (10a) as (10b) but not inversely.

(10) a. The elephant hits him.
b. The elephant hits himself.

In the model of this paper, this is easily understood. These children are in a phase where production of reflexives has been learnt, but where simulation in understanding with respect to reflexives has not yet taken effect. As pure concepts, the two pronouns can be described as (11).

(11) him
associated concept: a salient old referent
himself
associated concept: a commanding old referent in the same clause

And this predicts that normal pronouns will have reflexive interpretations: the commanding old referents are also highly salient.

The explanation of this effect in **?** rests on the assumption of a very specific constraint system and breaks down if this is replaced by a more plausible one. Another explanation by **?** rests on a complication of the already difficult and hard to motivate non-standard bidirectionality proposed by **?**. Both of these approaches were developed with the specific aim of solving the asymmetry and are difficult to motivate without it. Here it is a prediction of the model: there will be a delay between acquiring production of some part of syntax and the onset of production simulation for that part of syntax.

It is matched by the similar prediction of a delay between the onset of self-monitoring in production and the acquisition of understanding. There is some evidence for this: correct use of optional particles seems to very late (**?**). But a similar study on other phenomena for which self-monitoring seems plausible like word order freezing or optional case marking has to my knowledge not been carried out. The prediction would be that (12a) can be produced for (12c) up to some age and be understood as (12c) even longer before full adult blocking is reached.

(12) a. Mat ljubit doc
b. Mother loves daughter.
c. Daughter loves mother.

The explanation of **?** for the asymmetry is the late onset of bidirectionality due to resource limitations. This makes the problematic prediction that in old age a similar asymmetry would be observable. There are indeed observations that self-monitoring fails in old age, e.g. that pronouns get used more frequently when the antecedent is not easily accessible.

This prediction is not carried over in Bayesian interpretation. The inhibitory effects of simulation and automatic self-monitoring merely gets stronger and there is no reason why it should get weaker in old age. Non-automatised self-monitoring can indeed become weaker.

5 Conclusion

The paper presents a model of language interpretation (and production) which has a strong fit with looking at the brain as embodying Bayesian optimisation in perception, reasoning and other tasks (**?**, **?**, **?**).

The aim of this paper was to show that it leads to a natural temporisation and explanation of asymmetries in acquisition. It contrasts with the model assumed by **?** in having a stable architecture in which the bidirectional links become stronger when they start having a beneficial effect. It is moreover compatible with any reasonable syntactic description of the phenomenon.

The Bayesian approach to interpretation is a rich source of insights and the application to language learning is just an example. It was surprising to at least the author —though he should not have been surprised in the light of the work of Hobbs— that pragmatics comes out as probability maximation of the message (under the assumption that the speakers accommodate for perceptual bias).

Another surprise is that the Bayesian view comes out squarely on the side of those who advocated a production perspective in syntax. Early transformational grammar in which most of the action is in the mapping of deep structure (conceptual structure) to surface structure, generative semantics, functional grammar, systemic grammar, harmonic grammar and standard optimality theory are cases in point and each leads to a model of the probability with which a message is realised as a particular utterance. Harmonic grammar and stochastic optimality theory are merely the most sophisticated proposals by being the only ones that integrate probabilities.

The model also predicts problems in interpretation when the profile of $p(u|i)$ becomes too flat for the i parameter: for a particular beam search for an utterance there should preferably be one peak only for the candidate interpretations found. This will make interpretation more reliable. Morphology and/or regulated word order would be the devices that bring this about. So Bayesian interpretation predicts that there is syntax. Even if a language seems to lack any of these formal devices (e.g. **?**) there must still be strong preferences in production that can fill this functional role, unless speaker self-monitoring (i.e. the natural probabilities of the message) can take over entirely.

References

Blackmer, E. R. and Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39:173–194.

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*, 32:45–86.

Bresnan, J., Cueni, A., Nikitina, T., and Baayen, R. H. (2007). Predicting the Dative Alternation. *Cognitive Foundations of Interpretation*, pages 69–94.

Doya, K., Ishii, S., Pouget, A., and Rao, R. P. N., editors (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding.* MIT Press.

Friston, K. and Stephan, K. (2007). Free energy and the brain. *Synthese*, pages 417–458.

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.

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.

Jäger, G. (2003). Learning constraint sub-hierarchies. the Bidirectional Gradual Learning Algorithm. In Blutner, R. and Zeevat, H., editors, *Pragmatics and Optimality Theory.* Palgrave.

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.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, 14:41–104.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation.* MIT Press.

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

Noveck, I. A. (2000). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2):165–188.

Oaksford, M. and Chater, N. (2007). *Bayesian Rationality.* OUP.

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition*, 77:97–131.

Rizzolatti, G. and Arbib, M. (1998). Language within our grasp. *Trends in Neurosciences*, 21:188–194.

Shipra, J. B., Dingare, S., Christopher, Manning, D., Butt, M., and (editors, T. H. K. (2001). Soft constraints mirror hard constraints: Voice and person in english and lummi. In *Proceedings of the LFG 01 Conference. CSLI*, pages 13–32. CSLI Publications.

Smolensky, P. and Legendre, G. (2006). *The harmonic mind : from neural computation to optimality-theoretic grammar.* MIT Press.

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

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.