DOUBLE BIAS

Henk Zeevat

Abstract. The paper introduces a statistical notion of bidirectionality aimed at bridging the gap between synchronic and diachronic linguistic explanations.

1. Introduction

This paper presents a simple view on explanations of natural language use: a language user correctly uses this form in this situation because it is what others have used and would use in the same situation. The correctness criterion for the employment of a certain expression in a situation is outside the speaker and rests on the language game that is constituted by the other language use. And the same holds for the interpretation by the hearer of language use. The interpretation is correct not because the hearer has followed private rules for deriving the interpretation but because her reaction to the utterance is correct given that the others in the community would interpret the utterance in the same or closely related ways. In the final analysis, the explanation of why a language user like little Tom asks for "an egg" is nothing more than that he has heard others ask for eggs in that way and the reason that his mother gives him one is nothing else than that she has heard others ask for eggs in that way.

Computer technology has made it possible that usage is readily available for analysis and that it can be brought to bear on the technological problem of natural language analysis and generation, in its most radical forms without any other linguistic resources.

I want to defend a less radical view which has a strong

relation to optimality theory in its bidirectional conception (perhaps it is wrog to think of it as a different theory). In this view, both interpretation and generation are statistically biased and there is little else. One can however ask for the reasons why certain regularities are in the use and come up with synchronic and diachronic explanations very closely related to traditional linguistics or with sociological and psychological explanations relating to habit formation, preferences, or processing factors. Working out the effects of bias needs a theory of language structure and of its psychological realisation and it definitely pays off to use as much of linguistics and psycholinguistics as one can. But the claim is that once we have worked out the concept of in sufficient detail, we have a general theory of language structure (synchrony, generalisations over use) of language change (diachrony, the evolution of use) which fits into more general theories of human agents operating in societies.

2. $BIAS_{INT}$

An important constraint in interpretation is the principle that one selects the most plausible reading of a constituent given what one has: its linguistic context and the features of the constituent itself.

A constraint of this kind has been proposed by Mattausch Mattausch 2001 in the context of discourse processing (he calls it $CONSISTENT$) and it prefers readings that are plausible over readings that are not plausible.

Thus a reading of (1)

(1) John fell. Bill pushed him.

where Bill's pushing John caused John to fall is in accordance with the constraint and the reading where Bill's pushing follows John's fall goes against it. There are numerous interpretations of what plausibility here could mean (defaults, causal laws, etc.) but I adopt a probabilistic one here and throughout the rest of this paper. Here the probability relates to interpretation: if a pushing and a falling must be connected in discourse and the theme of the falling is the theme of the pushing than the pushing is the cause of the falling.

I call the probabilistic interpretation of the constraint $BIAS_{INT}$[1] and will try to defend a general version of it in terms of features. The generalised version says that if a linguistic construct is incomplete with respect to some feature than the incompleteness is resolved by filling in the most probable value given those features of the constituent and its context that are already known.

It is obvious why this is a sensible thing to do. Choosing an arbitrary value (with probability > 0) is just not as good as choosing the most probable one: it increases the chance that we are wrong. It makes no sense to gamble against the odds unless more can be gained by that. But the structure of the rewards is simple in this case. We either misunderstand or understand. It makes no sense therefore to gamble against the odds.

In Zeevat and Jaeger 2002[2], we explore the idea that $BIAS_{INT}$ can be used to give a functional/historical explanation of differential object marking. It can be shown by corpus studies on naturally occuring dialogue that high prominent subjects and low prominent objects are the statistical norm and it is obvious that in transitive sentences it is an essential task of the interpreter find out which is which. In the absence of subject markers and object markers, contextual factors (including expectations about the kind of argument of the particular verb), agreement and word order is all that remains. If word order in the language also marks other features, a certain amount of misunderstanding will be

[1] $BIAS_{INT}$is an abstract expression of the concept of Data Oriented Parsing (DOPBod and Scha 1997) by Scha and further developped by Bod and others. The idea there is to find the parse tree for a input sentence that has the highest probability given a corpus. This parse tree can be taken as input to an interpretation module and so can be seen as part of the interpretation process. For definitions, see Bod. Here it suffices to point out one difference: we want to take the best choice at choice points, i.e. we use linguistics to find the interpretational problems that need to be resolved. The level of trees is skipped, though it can be reconstructed. We are also not interested in probabilities except those that are relevant for the choices and think that these can be computed off-line given a set of classifiers. The principle $BIAS_{GEN}$ can be seen as a version of a still to be developed DOG (Data Oriented Generation). The bidirectional versions of these may be interesting from a theoretical point of view.

[2] Closely related conclusions are reached by Haspelmath 1999

generated. The adoption of transitive subject markers or object markers will reduce the confusion and liberate the word order dimension and so increase the overall efficiency of the language. The accusative pattern arises if just an object marker is adopted, the ergative pattern if a marker is adopted for transitive subjects and split ergative results from having both subject and object markers. The extension of the marking is a function of the other marking possibilities that are available. Further work is necessary to establish this kind of explanation in a formal framework like e.g. evolutionary game theory.

Another explanation along the same lines is the explanation of the following universal. While it is possible and in fact sufficient for psychological attitudes that what the subject is glad about or what she regrets is a belief of that subject - witness (2)- that is not its presupposition: that is that the complement is a fact (at that point of the conversation).

(2) John thinks that Mary is coming to the party and he is
glad that she is coming.

We have to assume that a predicate like *be glad that* can only take complements to propositions that are given and that are believed by the subject. Now while it is possible that the proposition is given as as a belief of the subject, in the overwhelming majority of cases the proposition is given in the common ground as a fact and a simple bridging inference (that the subject is aware of the fact) is required for establishing this. The probability that the speaker assumes the truth of the proposition and the corresponding belief of the subject therefore far outweighs the probability that the speaker only assumes the belief of the subject. This makes the accommodation (a repair strategy) follow the observed pattern.

$BIAS_{INT}$ automatically covers fully determined choices. In that case the probability is just 1. It also covers other cases. An irrelevant interpretation can be discarded because of the unlikelyhood of irrelevant messages being offered, but we do not always pick the most relevant reading. The German word *Esel* has a reading where it means "male donkey" which is always at least as relevant as the more common meaning "donkey". But the extra relevance has to override the overwhelming probability of the default reading, and seems to do that only in the presence of a formal contrast, as in (3).

(3) Das ist kein Esel, sondern eine Eselin.
That is not a donkey but a she-donkey.

If a feature is missing in a partially interpreted linguistic expression L and there are choices $C_1 \ldots C_n$ with probabilities $p_1 \ldots p_n$ for $C_1 \ldots C_n$ in the context c, for the expression L choose the C_i that has the highest p_i. What do we do if there is rough parity of probability? In that case, both options will have to pass. This is not unrealistic, since overt ambiguity does happen. A perhaps preferable option is to work with sets of interpretations. This option seems almost necessary in the intermediate stages.

There are two unclarities here. First of all, in order to have probabilities we need a classification of $L + c$ in terms of a finite set of features. It is reasonable to assume that certain features do not have a bearing on the current choice and that consequently we can limit ourselves to a subclass for any specific feature. What these features are is decidable if we have selected the total class.

It is clearer what are the incompletenesses that have to be decided. A linguistic expression comes in as a sound pattern which is mapped to a sequence of words that acquire grammatical roles and semantic roles and identities with elements in the context. Each of these acquisitions is a feature of the relevant kind. $BIAS_{INT}$ should then regulate the assignment of words to sounds, the assignments of meanings to those words, the dependencies between the words, the anaphoric relationships, the discourse relations, the information structure etc.

$BIAS_{INT}$ also disfavours interpretations that are not consistent. Such interpretations are normally not correct, though they are possible if they are explicitly marked as such, as in corrections. So it may be argued that most of the pragmatic principles that one would like to assume in a full-fledged interpretation theory actually will fall out.

I have suggested (unfortunately not proven) that $BIAS_{INT}$ can take the role of all the defeasible interpretation principles that I proposed in Zeevat 2001. There we have

$FAITH_{INT} > CONSISTENT > *INVENT > RELEVANCE$

$BIAS_{INT}$ replaces (by the way it is formulated) most of the work of *INVENT, the principle that enforces the resolution of pronouns and other devices that refer to features of the context. But it also enforces relevance to the extent that usage contains it. (I favour a formulation of $RELEVANCE$ where the utterance addresses a goal or goals (these can be questions) that are activated in the context). And it militates against utterances that are not consistent with the common ground, if it is the case that these are sufficiently rare (only the cases that are not marked as correcting the common ground count here). I do not propose to derive the first principle from $BIAS_{INT}$, though it could be derived in that way. It says that one should take all the formal aspects of the utterance into account and so provides an explanation of why (4)

(4) Pete is coming after all.

can be taken to replace information in the common ground or why (5)

(5) John fell. But not because Bill pushed him.

does not allow the biased interpretation we had before. Though the principle can be derived from $BIAS_{INT}$, it is not clear to me how its non-violability (i.e. its position in the hierarchy) follows from the argument. That is better understood from the combination of $BIAS_{GEN}$ and bidirectionality.

3. $BIAS_{GEN}$

We can add a second constraint, BIAS in the other direction, $BIAS_{GEN}$. Here the speaker adds features to her linguistic construct on the basis of the probabiblities in usage. If normally a formal feature is added to a construct with a certain profile in a certain context, then it is done now again, if the speaker is sufficiently part of the usage. The notion is the same but this time the semantic features are given (with the context) but the formal features of the utterance need to be determined. What features need to be determined is not a difficult question: a full utterance needs to emerge, with words, morphology, intonation and word order. (The characterisation of the given features is more difficult, but we can take our lead from interpretation theories). There is a clearly functional demand here: the utterance that emerges should be understandable to the interpreter. Why certain features need to be present (when they do not come from the requirements of putting meaning into sounds) should have its explanation there or in history.

What does the principle explain? An old puzzle is the question why idiomatic expressions win. German and English are sufficiently alike to allow the formation of equivalents of each other's way of asking the time: (6)

(6) How late is it?

or (7)

(7) Welche Zeit ist es?

. Yet, these expressions are nearly incomprehensible to normal speakers of either languages. It is clear that in the circumstances, there is a very strong preference for using the standard expression and not something that needs to be generated and interpreted by a compositional process. $BIAS_{GEN}$ explains why it is used almost exclusively. Much the same holds for other instances of idiom, e.g. the peculiarities of the use of prepositions in languages like English. Only one use is right (e.g. *in his office* and not *on his office*, as in Dutch), but the meanings of the prepositions seem sufficiently elastic to allow interpreters to guess the intended relation.

This generalises to all cases where languages seem to have taken an arbitrary decision. In that case, there is a 100% bias. If the situation is best described with defeasible rules as in optimality theory, these again will have bias working with them. $BIAS_{GEN}$ clearly militates towards compliance with the constraints, if the appropriate triggering conditions are generated.

4. Bidirectionality

The full power of double bias is reached when we condition $BIAS_{INT}$ and $BIAS_{GEN}$ on each other. We do not want to have that the biased expression α of an input x does not receive the biased interpretation x. That would defeat the purpose of verbal communication. Similarly, we do want to assign a biased interpretation x to α if α is not the biased expression of x. In that case, the interpretator fails to understand why the utterer of the expression has chosen that expression and the situation does not fall under the Gricean concept of non-natural meaning: the intention of the speaker is not grasped. If the biased interpretation is adopted, conversational implicatures in accordance with the cooperation principle arise to provide further explanations of the speaker's intention with his non-standard utterance. The problem is not of the same order as the failure of the first kind of bidirectionality: it is always possible to accept the biased interpretation and assume that the speaker is making a performance mistake due to fatigue, illness, linguistic incompetence and the like. Or is just being uncooperative.

We can turn this into a formal definition (8).

(8) **Bidirectionality**
α is an optimal expression for x iff
α is the most probable expression of x and x is the most
probable interpretation of α
iff x is an optimal interpretation of α

But what should people do if there there is no optimal expression or optimal interpretation. They then have an alternative:

Speaker: I choose α as an expression for x, because in that way I maximise the chance that the hearer will understand my message as meaning x.

The speaker is not necessarily choosing the form α because it is the form that gets assigned the highest probability as an expression of x, because x might not be the interpretation that has the highest probability of being chosen as the interpretation for α. On the other hand, moving away too far from the probablistic generation optimum runs foul of the hearer strategy. The speaker will choose the item with the highest probability as an expression that maximises the chance of being understood by the hearer.

The hearer is also not blindly taking the least marked interpretation, but taking into account how that maximum should be expressed. If α is not the optimal form for x, then that counts against the interpretation.

Hearer: I choose x as the interpretation of α because the chance is maximal that the speaker has chosen α as an expression of x.

Both strategies are common ground between speaker and hearer, so that the speaker has to take into account the hearer strategy and the hearer the speaker strategy. This forces them into a compromise, the hearer will maximise the recognisability of her message without making it a too marked way of expression: that will put the hearer of target.

The net result is a simultaneous maximisation of the probabilities on both sides. The sum would allow one side to be very small, provided the other side does well. So the product of the two probabilities is a good approximation of the equilibrium. So we can reformulate the two strategies as.

Speaker: I choose α as an expression for x, because in that way I maximise that the chance that the hearer recognises my intention to express x. (I maximise the product of the probability that x is expressed as α and the probability that α is interpreted as x).

Hearer: I choose x as the interpretation of α because that maximises the chance that that is what the speaker intended. (I maximise the product of the probability that x is expressed as α and the probability that α is interpreted as x).

Under this definition the natural definition of the quality of a form-meaning pair is the probability we get by taking the products of the two probabilities. Choosing optimal interpretations and generations then converges towards choosing for interpretations x for α such that the the probability that x is the correct interpretation of α multiplied with the probability that α is the realisation of x is maximal.

A crucial observation is that by following bidirectional bias we do not reproduce use, as we would when we would just be following $BIAS_{INT}$ and $BIAS_{GEN}$. If there is a mismatch (α is more probable for x, but y is more probable for α), bidirectionality makes it the case that α is less often chosen for x and that y is less often chosen for α (if y is not most probably realised as α). Matches increase both probabilities, mismatches decrease them. Decreasing probabilities can lead to the emergence of alternatives. If there is a proper match, bidirectional bias increases the probabilities in both directions and can block other equally probable interpretations and generations.

There is a direct application to diachrony here. If we follow Croft Croft 2000 in accepting that the items that linguistic evolution reproduces are utterances and we accept the above account, we obtain almost directly the elements of a theory of evolution.

First of all, the speaker and hearer do not have access to the real probabilities. They have to make estimates based on a finite sample that may not be representative. Especially for low frequencies the estimates can be very unreliable. Speaker and hearer have to allow for deviation from the norm given by the actual probabilities. Also the estimation of the probabilities for complex expressions/inputs on the basis of the proabilities of their parts may be very inaccurate. This predicts variation even when we start with the same actual use.

Every decision to produce an utterance is an attempt to maximise the product. Selection between variants is therefore part and parcel of the theory: it is the heart of a theory of blocking, of grammatical correctness, of interpretation etc. Expressions can become bad because they lose out in frequency (salience) to other expressions or because they are less well understood than others.

It may be true that even innovation (the invention of new linguistic devices) can be brought under this perspective. Assume that the input is sometimes only realisable by expressions which are blocked, because they lead to interpretations that are different from the input. This forces expressive innovations. These can be either marked new forms that are still interpretable. Bidirectionality predicts that these forms are special for the interpreter: she has to explain why the speaker has not chosen the most conventional means of expressing herself. They can also be metaphorical uses. Here the message is conveyed by an unsuitable predicate or description, unsuitable in the sense that it could never apply to what it is supposed to apply. Again, the hearer will have to make sense of why the speaker chose this device and is this time guided by the words chosen by the speaker and the images and associations they convey.

The concepts of reproduction, variation, innovation and selection together form a concept of evolution.

Following Blutner 2001, we can make the definition recursive by replacing *biased* by *optimal* in the definition of the two sets. I am not sure this brings very much in the current setting. The reason for thinking this is that we can already explain two tendencies in natural languages for which the recursive version seemed necessary.

The first is the relationship between word length and the frequency of the words. The generalisation is that high frequent words tend to be shorter than low frequent ones. This is a tendency. In the first days of the advent of the television, the frequency of the word *television* must have risen dramatically, but it must have taken some time before the abbreviation *teevee* took over. There is nothing contradictory in high frequent but long words. The opposite also exists: *eg* in Dutch is an obsolete agricultural instrument and the word must have been much more frequent than it is today where -outside farming communities- its use is restricted to the language of the crossword puzzle.

How does the explanation go? Low frequent words need more phonetic characteristics for their recognition since they have a low activation level. High frequent words can do with less. Length correlates with the number of phonetic characteristics. The moment there are different words with different lengths for the same concept and a high frequency exists for the two words taken together, we need to explain why the shorter word will win, i.e. will be reproduced more often than its longer counterpart.

Next to double bias, we here need two extra factors. One is the well-established relationship between frequency and ease of recognition. This explains why short forms for frequent words do not lead to extra recognition problems. (Speakers should avoid these, by bidirectionality.) The second factor is more puzzling and harder to provide with an empirical underpinning: a natural preference for short words over equivalent longer ones under high frequency. Laziness and reduction of effort do not seem of primary importance: it does not take that much extra effort to pronounce the longer words and certainly one does not achieve enough of an energy gain to score better in natural selection. I will assume that it is tedious to use a long word frequently. Ease of recognition allows shorter forms, tedium prefers them.

Examples are: *OT* for *optimality theory* (among *cognoscenti*), $fiets$ for *vélocipède*, *phone* for *telephone*, *cd* for *compact disk*, etc. We have to assume the emergence of small communities within which the shorter word is dominant and spread under preference in conditions of unclarity about which convention obtains in a certain conversational setting. The possibility of local convention explains why such small communities can emerge.

If this is correct, it also gives a partial account of the iconicity effect. Here we have to assume that we have two expressions with the same meaning, where one becomes dominant. By the considerations above, this would be the shorter one. If the longer one does not disappear, it will be the marked way of expressing the same concept and lead to the need of the interpreter to find an additional explanation for the speaker's unusual choice of expression. Conversely, if the set of instances of the concepts falls apart into a set of standard cases and into a much smaller set of non standard cases, a referential occurrence of the unmarked expression will lead to the conclusion that the reference is not to a non-standard case. A natural explanation for the use of the marked expression is therefore that one is dealing with a non-standard case. An implicature is generated. If the speaker says (9),

(9) Black Bart caused the sheriff to die and he did so in a very normal way.

she is not contradicting herself, though she does not give away why she chose to use the marked expression.

Bidirectionality is also central to the explanation in Zeevat&Jaeger 2002. $BIAS_{INT}$ suffices for showing that absence of object or subject marking can lead to confusion. High prominent objects are misrecognised as subjects, low prominent subjects as objects. But the changes that are necessary to remedy the situation cannot be explained from $BIAS_{INT}$ alone. The force behind history that is at work here is the first half of **Bidirectionality** which outlaws wrong interpretations. Generation should not just be oriented at minimising markedness, but also at understandability. Devices promoting understanding will naturally creep into usage and thereby influence the working of $BIAS_{GEN}$ and $BIAS_{INT}$. In the paper, we show how new object marking changes the bias for interpreting unmarked high prominent objects and thus further increases the need to mark high prominent objects. $BIAS_{GEN}$ picks up the same usage and can promote marking by making it obligatory by making it a rule rather than a way of helping interpretation.

References

Blutner, R.: 2001, Some aspects of optimality in natural language interpretation, *Journal of Semantics* 17(3), 189–216

Bod, R. and Scha, R.: 1997, Data-oriented language processing, in S. Young and G. Bloothooft (eds.), *Corpus-Based Methods in Language and Speech Processing*, pp 137–173, Kluwer Academic Publishers, Boston

Croft, W.: 2000, *Explaining language change : an evolutionary approach*, Longman, Harlow, England & New York

Haspelmath, M.: 1999, Optimality and diachronic adaptation, *Zeitschrift für Sprachwissenschaft*

Mattausch, J.: 2001, *On Optimization in Discourse Generation*, ILLC report MoL-2001-04, MsC Thesis, University of Amsterdam

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

Zeevat, H. and Jaeger, G.: 2002, A statistical reinterpretation of harmonic alignment, in D. de Jongh, M. Nilsenová, and H. Zeevat (eds.), *Proceedings of the 4th Tblisi Symposium on Logic, Language and Linguistics*, ILLC, Amsterdam, ICLC, Tblisi, Amsterdam