Towards a Probabilistic Version of Bidirectional OT Syntax and Semantics

Kees van Deemter
ITRI, University of Brighton
Kees.van.Deemter@itri.brighton.ac.uk

This paper argues that a purely probabilistic version of bidirectional Optimality-Theoretic syntax and semantics (Blutner 2000, Zeevat 2000) can make some significant contributions to the computational processing of natural language. The ideas outlined in this paper apply to interpretation as well as generation, but particular attention will be given to the question how bidirectionality can be put to use in Natural Language Generation. The paper sketches a number of different possible approaches, but one of them will be described in more detail, at the heart of which lies an empirically oriented notion of *vicious ambiguity*.

1 Optimality Theory (OT)

Optimality Theory (OT) is a family of theories, originally inspired by work on neural networks, which have come to play an important role in linguistics, starting with phonology and morphology (Prince and Smolensky 1991), while some more recent versions cover syntax, (e.g., Bresnan (in press), Frank et al.1998, Kuhn 2003) semantics, and pragmatics (Blutner 1998, 2000; Beaver, in press). OT theories differ in their details, but they share a sense that the degree of 'harmony' between inputs and outputs of a linguistic module depends on a number of factors. OT is especially suitable for the modelling of phenomena where different factors (modelled as constraints) point in different directions. Given a certain input, for example, one candidate output may violate a constraint c_1, while another candidate violates constraint c_2. If this happens, a balance may be struck by declaring one of the constraints, for example, c_1, more important than the other, causing the output that violates c_1 to be judged 'less optimal' than the other one *ceteris paribus* (i.e., other violations being equal). Constraints induce a ranking of the different outputs depending, roughly speaking, on which constraints they violate and how often.

In this paper, we bring ideas from OT to bear on computational interpretation and generation, focussing on the question how a Natural Language

Generation (NLG) program can select the 'best' formulation from among a set of candidate formulations, each of which is grammatically correct (i.e., they all correctly express the meaning of a given input). In particular, we will explore how ideas from *bidirectional* OT – a version of OT that has become particularly popular among students of the interface between syntax and semantics/pragmatics – can contribute to this task. To allow bidirectional OT to play this role, we will propose a number of changes in the way in which bidirectional OT is usually conceived. Some of these changes are caused by the fact that the ranking of inputs and outputs will no longer be derived from constraints, but directly from the frequencies in a corpus: The likelihood of a pairing between a given type of Content and a given type of Form (cf., Briscoe and Carrol 2002, Johnson et al. 2002) will induce a ranking of all the Contents associated with a given Form, and of all the Forms associated with a given Content.

Further modifications to the dominant version of OT syntax/semantics will be caused by the assumption that *other* theories will tell us which Forms can express which Contents, while the OT-inspired framework will tell us only which of these is the best expression of a given Content. The main effect of these adaptations will be to shed light on the avoidance of ambiguity in generated text, in a setting where other properties of a text are also taken into account, such as its conciseness. We will also, rather more briefly, explore a reversal of this setup, in which a probabilistic version of OT is used to select the most plausible interpretation of an utterance from among a set of grammatically possible interpretations.

The perspective on OT that we will be proposing appears to be novel, even though OT and corpus-based probability have been coupled before. In Boersma (1997), for example, unlike our proposal, constraints continue to be central (although their ranking varies stochastically, to account for variation in language production). More recently, Zeevat (2002) has started to explore statistical, corpus-based methods for OT semantics and pragmatics but his paper focusses on applying NLP to OT, rather than the other way round. A number of other approaches relating NLP and OT will be discussed in what follows.

The essentials of this programmatic paper can be made concrete in many different ways. We feel, however, that the paper would have lacked concreteness unless at least one of these different formalisations was presented in some detail, and this will be done in sections 5-9. Some variants of the approach of sections 5-9 will be sketched in section 11.

2 OT for NLP

OT explains why certain Contents ought to be expressed in certain ways, and why certain Forms ought to be interpreted in certain ways. Typically, OT does not assume the existence of a grammar separate from OT. To consider a classic example, *'John washes him'* does not normally mean that John washes himself; in OT, this interpretation is 'blocked' because it would have been better (i.e., more optimal) to have expressed it as *'John washes himself'*. This replaces other mechanisms, such as the binding rules (Chomsky 1981), as an account of certain facts about anaphora. Although OT might enable one to account for a larger class of observations (e.g., Burzio, 1989), we will explore a different way in which optimality-theoretic notions can contribute to NLP. From a generative perspective, the point is the following. The key challenge for NLG is to find the *best* way of expressing a given Content from among all *correct* ways of expressing it: it requires finding one expression, among all the correct ones, that can lay a reasonable claim to being optimal. Ultimately, this choice should depend on the linguistic context of the utterance[1], but it will be convenient to disregard this for now, by pretending that the problem is simply one of mapping a Content to a Form. In the following section we will discuss how some types of nonoptimality in generated texts can be avoided or repaired.

For future reference, we will call the perspective on OT outlined in the previous paragraph the *selection* perspective on OT, to distinguish it from the dominant 'grammatical' perspective, (in which OT is a mechanism for determining grammaticality). This perspective can apply to interpretation too, since it is a key challenge for Natural Language Understanding to select the most plausible intended interpretation from amongst all interpretations that are grammatically possible. The selection perspective on OT has been brought to bear on the problem of parse ranking in such works as Harrison (1988), Frank et al. (1998), Butt et al. (1999), but the approach to Natural Language Understanding that we will advocate in section 8 will differ from these earlier endeavours by being bidirectional and corpus based.

[1] See, for example Toutanova et al (2002) for the use of 'contextualised' versions of Probabilistic Context Free Grammars (in particular PCFG-A's).

3 Controlling ambiguity in NLG

One aspect of controlling the choice between different wordings in NLG has received more attention than most: the problem of controlling ambiguity. Usually,[2] this has taken the Gricean form of trying to avoid certain types of ambiguity. Guaranteeing uniqueness of *reference* is, of course, the linchpin of most algorithms in the generation of referring expressions (Dale and Reiter 1995; also van Deemter 2002), where properties are accumulated until the resulting description applies to only one referent (i.e., until the intersection of the extensions of the accumulated properties is a singleton set). To an extent, this might be seen as a metaphor for language generation in general, if one views the task of a speaker as an effort to identify the intended interpretation, at the expense of all 'distractor' interpretations.

Avoidance of ambiguity has started to be an issue in NLG, because NLG systems are acquiring larger and larger linguistic coverage, making the choice between different expressions more and more difficult. Controlling ambiguity is of special importance in those systems that try to generate texts that are clear as well as 'natural'. (Examples include user manuals, medical information sheets, etc., which are aimed at a non-specialist audience while also conveying detailed and important information in ordinary language.) A representative example of an NLG architecture that tries to avoid ambiguity is the work by Inui and colleagues, who designed a 'reflective' architecture structured around a revision loop (Inui et al. 1992). A system incorporating such an architecture is called reflective because it 'reflects' on its own output.[3] Reduced to its core, their approach can be schematised as follows. One assumes that the set of syntactic revision operators is finite and that they are ordered in a list.

```
GENERATE & REVISE:
   - Draft := A sentence generated from the input
   - Repeat
       - Choose revision operator R from the list
       - If R is applicable then Draft := Revised(Draft)
     Until all revision operators have been attempted
   - Return Draft
```

The algorithm visits each member of the list of revisions, applying a revision if it is applicable (in which case we will assume that it is guaranteed to lead to an improvement). Revision operators were designed to correct two types of problems in generated Japanese text: excessive complexity and excessive

[2] *Pace* Neumann and van Noord (1992), who seek to *respect* the ambiguities found in an expression of a source language when translating it into a target language.

[3] Reflection has also be called *monitoring* (e.g., Neumann 1994, 1998), a term which is properly reminiscent of the various (self-)monitoring processes that psycholinguists have argued for (e.g., Levelt 1989, chapter 12).

ambiguity. For an English example, suppose one wants to say that all old men and all old women are entitled to certain benefits:

$$\forall x((Man(x) \vee Woman(x)) \rightarrow (Old(x) \rightarrow Ben(x))),$$

and suppose an initial Draft of the form

'All old men and women are entitled to benefits'

is generated, which is ambiguous with respect to the scope of '*old'*. Inui et al. argue that ambiguities of this kind can be difficult to foresee, making some sort of revision strategy inevitable. To remove the ambiguity from our example sentence, the Draft may be revised into *'All old men and old women are entitled to benefits'*. The idea of revision has proven to be a seminal one whose potential applications are numerous. (For example, one might imagine a language generation system that analyses its own output to detect and eliminate unwanted implicatures.) Moreover, as we will see later, the attempt to tackle ambiguity and complexity through one and the same method has proven to be farsighted. Variants on the idea of a revision architecture have been proposed by a number of authors (Meteer and McDonald 1986, Yazdani 1987, Robin 1994, Callaway 2003). Of particular interest for us is the approach in Neumann (1994, 1998), which refines Inui's approach to avoidance of ambiguity by bringing computational *grammars* to bear, allowing the system to determine precisely (1) whether a given string is ambiguous (given a grammar), and (2) what all the alternative wordings of the input are (given a grammar). As is natural, Neumann uses the same, reversible, grammar in both cases and this has additional advantages in terms of psychological plausibility as well as software management. Speaking very generally, it is this type of approach to the problem that we will advocate, but we will add a number of new considerations to it. Foremost among these is an assessment of the relative *plausibility* of a given parse (or interpretation). The reason for this move is that, when grammars become large, they assign multiple parses to more or less *every* sentence (even when spurious ambiguities are put aside), many of which tend to be highly unlikely. This, of course, causes avoidance of all ambiguity to be an impossibility (e.g., Abney 1996). It is here that statistical parsing can be of help, as we will see in section 5, since it allows us to estimate the likelihood of a parse and, thereby, to separate the sheep from the goats.

4 Bidirectional OT

But first, let us reflect on what has been suggested by such authors as Inui: for a given grammatically correct pairing of a Content C (i.e., a semantic representation of some sort) and a Form F (e.g., a sentence), they wonder whether improvements are possible. Improvements of two kinds are considered: those that couple C with a more desirable (e.g., shorter) Form, and those that couple C with a Form, say F', that is less ambiguous than F. But what does it mean to say that F' is less ambiguous than F? Evidently, this must have something to do with *other* Contents that F and F' can express. In other words, the reflective approach to avoidance of ambiguity can be construed as looking in two directions: forward, to all the Forms that can express a given Content, and backwards, to all the Contents that can be expressed by a given Form.

To people familiar with work in Optimality Theory, this story has a familiar ring to it, since OT has also started looking in both directions (i.e., generation and interpretation) at the same time, based on the 'grammatical' perpective (section 2):

> 1. When trying to find a suitable interpretation of a given Form, other Forms must be taken into account as well (e.g., Hendriks and de Hoop 2001.) An example is the interpretation of *'He washes him'* (section 2), where one interpretation is 'blocked' by a Form that expresses it more appropriately: *'He washes himself'*. This (basically Gricean) approach has become standard in much of OT.
> 2. When generating from a given Content, other Contents might also be taken into account. Like in interpretation, the key question is whether a given Content is more appropriate (in relation to a given Form) than another (Suggested by Zeevat 2000, p.255).

When both directions are combined, we speak of a *bidirectional* version of OT. In the above-mentioned earlier proposals, appropriateness is measured by checking, for a number of separately specified *constraints*, which of them is violated by a given pair of a Content and a Form. Most constraints that have been proposed focus on the appropriateness of a given Form (e.g., Hendriks and de Hoop 2001), but more recently, constraints have also been proposed that focus on the appropriateness of a given Content (Zeevat 2000). For any Content C and any Form F, the inspection of constraints leads to orderings of the sets Forms_C and Contents_F.

$$\mathcal{E}(C,F) \Leftrightarrow F \text{ is a legitimate expression of } C$$

$$\begin{aligned} \text{Forms}_C &= \{G : \mathcal{E}(C,G)\} \\ \text{Contents}_F &= \{D : \mathcal{E}(D,F)\} \end{aligned}$$

We assume that the relation $\mathcal{E}(C,F)$ is given, possibly specified in a grammar along the lines of Copestake and Flickinger (2000), for example. The notation $(C,F) \prec (C,F')$ (which could be read as (C,F) is 'less harmonic' than (C,F')) means that (C,F) incurs more violations than (C,F') of the highest-ranking constraint for which the two pairs differ.[4] This is sometimes written,

[4] This somewhat surprising arrangement, by which one constraint violation cannot be

in the reverse order, as $(C, F') < (C, F)$ (e.g., Jäger 2002). For reasons that will become apparent later we follow Blutner, writing $(C, F) \prec (C, F')$, reflecting the fact that F occurs less frequently than F'. The meaning of $(C, F) \prec (C', F)$ is exactly analogous. Note that the relation $\preceq$ induced by this method is reflexive and transitive (i.e., it is a *preorder*) and neither asymmetric nor antisymmetric (so it is no *partial* order).

Bidirectional OT usually works by requiring the combination of a Form F and a Content C to fulfil a requirement of *superoptimality.* Superoptimality can be formulated in different ways, as we shall see. The best known version takes its origin in Blutner (2000). Paraphrasing Jäger's (2002) formulation of Blutner (2000) slightly, 'strong' superoptimality can be formulated as follows:[5]

> **Strong superoptimality (original version)** Let $\mathcal{E}(C, F)$. Then (C, F) is superoptimal iff there is *neither* a Form F' such that $\mathcal{E}(C, F')$ and $(C, F) \prec (C, F')$ *nor* a Content C' such that $\mathcal{E}(C', F)$ and $(C, F) \prec (C', F)$.

In other words: for (C, F) to be superoptimal, C has to be optimal with respect to F, and F has to be optimal with respect to C. Strong superoptimality in the style of Blutner is an intuitively appealing notion closely related to that of a *Nash equilibrium* in Game Theory (Dekker and Van Rooy 2000). To use a comparison proposed by Eisner (2002), Contents may be likened to children and Forms to the cloths that they might wear. In this juvenile domain, $\mathcal{E}(C, F)$ means that F *fits* C. Different children look best in different cloths, even if the same cloths fit all of them. Thus, $(C, F) \prec (C, F')$ means that F' looks better on C than F, while $(C, F) \prec (C', F)$ means that C' looks better in F than C. In this domain, a combination (C, F) is superoptimal if F is the best possible outfit for C while C looks better in F than any other child. It is easy to see that this makes (C, F) a highly suitable combination.

It is possible to apply these ideas from bidirectional OT directly to NLP (using previously-proposed constraints and the orderings of Contents/Forms induced by them to select the best Form for a given Content, or the best

offset even by multiple violations of lower-ranked constraints has been argued to be empirically appropriate for OT's original applications (e.g., Prince and Smolensky 1997). Whether it is also appropriate for OT semantics is a question that need not detain us here. See Power et al. (2003) for a different approach, based on constraint satisfaction.

[5] *Weak* superoptimality was designed to account for partial blocking, aiming to explain why unwieldy expressions like *'Black Bart caused the sheriff to die'* are sometimes preferred over apparently preferable ones (*'Black Bart killed the sheriff'*). Weak superoptimality appears to be irrelevant for our purposes.

Content for a given Form) and readers will have little difficulty working out the details for themselves after reading section 6. (See also section 11.2.) This would, however, probably not lead to a practically very useful result, because the constraints that have so far been forthcoming from the OT syntax and semantics community, although interesting for their potential explanatory value, are not finegrained enough to distinguish, for example, among all the Forms of the language that a generator can associate with a given Content. The problem may be clearest for the 'semantic' constraints, which distinguish between different Contents. Zeevat (2000), for example, proposes only six such constraints (of which he argues that they form a natural class) all of which are formulated at a high level of abstraction. One of them, for example, forbids Contents that are inconsistent with mutual knowledge, (Zeevat 2000, p.251). It is doubtful that such constraints could always distingish between the different interpretations of *'John saw the man with the telescope on the hill'*, for example, even if detection of inconsistencies was always computationally feasible.

For these reasons, we will explore a different approach. We will take Bidirectional OT as a starting point, but we will interpret *optimality* in a corpus-based, probabilistic way, inspired by methods from statistical NLP. We will propose modifications of OT where they become necessary as a result of this re-interpretation, and where they are dictated by the *selection* perspective on OT (section 2). The result will be an approach that shares the philosophy of bidirectionality with OT, while deviating from it in other ways. In particular, we will not use the notion of an OT-style *constraint*, deriving the ordering of Content/Form pairs from a language corpus instead.

5 Statistical NLP

The basic idea that we will explore here is to use frequencies in a syntactically and semantically annotated corpus to establish the likelihood that a given type of Content is expressed via a given type of Form, and the other way around. In OT semantics, *Forms* are syntactic constituents, but the nature of *Contents* tends to be left unspecified (Blutner 2000, Zeevat 2000, Jäger 2002). NLP systems are, of course, forced to represent Contents as concrete formulae of a formal language and most NLP systems use a variety of formal languages at different stages of processing. An NLG system, for example, may start out with something as unstructured as a set of atomic formulae, which are aggregated and modified during later stages of processing (e.g., Reiter and Dale 2000, section 3.3.6). The questions that we are asking in this paper could be asked of each of these stages of representation, but we shall focus on one of the later stages, where some of the sentential structure has already been determined. Wide-coverage tools for mapping between semantic representations and sentences are starting to be a realistic prospect. Of particular interest, to us, are *statistical parsers*, which estimate the likelihood, for a given sentence, that it has a given parse (e.g., Charniak 2000). Parsers of this kind are becoming increasingly accurate and have recently started to include a level of semantic analysis in the form of Minimal Recursion Semantics (MRS) formulas (Briscoe and Carroll 2002, Toutanova et al. 2002).[6] They allow us, for example, to predict the likelihood of each of the two interpretations of an NP like *'old men and women'* (i.e., depending on whether *'men and women'* is a separate constituent). Typically, such a prediction is based on frequencies in an annotated corpus.

Suppose we are interested in interpretations of the string F = *'old men and women'*. If we are incredibly lucky then this string will occur many times in the corpus, for example x times annotated with an interpretation C_1 that corresponds with the bracketing (old (men and women)), and y times with the interpretation C_2 = ((old men) and women), and z times with some other interpretation. Then the probability, given the Form F, of C_1 equals $x/x+y+z$, while that of C_2 equals $y/x+y+z$. In most cases, we will not be so lucky, and the string F might not occur in the corpus even a single time. The standard trick, in such cases, is to 'back off', that is, to look at increasingly more general patterns of which F is an instantiation, until sufficiently many occurrences are found. One might back off, for example, to the pattern *old + AnimateNoun1 + and + AnimateNoun2*, or even further, to *Adj + Noun1 + and + Noun2*. Conditional probabilities are then calculated as described earlier.

Annotating a sufficiently large corpus with all the relevant information (e.g., Kingsbury et al. 2002, Toutanova 2002) is a time-consuming and error-prone activity, and the details of the annotation scheme matter. For example, if 'children' and 'women' share all their annotation features (e.g., they are both animate nouns) then the predicted probabilities involving *'old men and children'* are the same as those for *'old men and women'*, which would lead to less than optimal results, because 'old' is a less probable modifier of 'children' than of 'women'. Nonetheless, statistical parsers are beginning to achieve very respectable results, measured in terms of the percentage of sentences for which the parse ranked highest by the parser is the one that annotators have designated as the correct one (e.g., Johnson et al. 2000, Toutanova et al. 2002). Similar methods, involving statistical 'structured' (i.e., syntactic) language models, are now used with success in speech recog-

[6] MRS is used here as an example because its underspecification mechanism allows the authors to record semantic information at the level of a (complete or incomplete) predicate-argument structure, which is compatible with the grammatical relations that are derived from the syntactic derivation trees (Briscoe and Carroll 2002, section 2.7).

nition (Chelba and Jelinek 2000). Moreover, the connection between the corpus-based probability of a given type of syntactic analysis and its probability in the mind of a human has started to be established in recent years.[7] Here, we will not delve into the details of this issue but take a minor leap of faith and assume that, at some point in the future, statistical parsing will offer a reasonable approximation of the likelihood that a given interpretation is the one intended by a speaker/writer. The perspective can also be reversed, using a corpus to assess the likelihood of a *Form* F given a *Content* C, and building it into a statistical *generator*.

Henceforth, we will be assuming that probabilities are derived from frequencies in corpora, and that the probability of a Form will be conditional on a given (type of) Content and the other way round. Various simplifications are possible, however. For example, one could disregard the Content C entirely (cf., the distinction between markedness and faithfulness constraints in OT, Kager 1999) and, for example, assess the probability of a Form using n-grams (Knight and Langkilde 1998). Going a step further, probabilities do not necessarily have to be derived from a corpus. For example, one might simply stipulate the probability of a Form to be inversely proportional to its length.

Let us introduce notations for the two different types of probabilities. Probabilities can be assigned at different levels of granularity (sentences, clauses, phrases, etc.) but, in our examples, we will mainly use sentences.[8]

$p(C|F)$ = given a Form F, the probability
that F is interpreted as C.
$p(F|C)$ = given a Content C, the probability
that C is expressed as F.

Probabilities of the second variety are of obvious relevance to NLG: if a Content C can be expressed through Forms $F_1, ..., F_n$ then we compare $p(F_1|C), ..., p(F_n|C)$; if $p(F_i|C)$ is largest then, *ceteris paribus*, F_i is the best expression. Ties (i.e., cases where $p(F_i|C) = p(F_j|C)$) may occur, and in such cases the choice becomes arbitrary.

But there is another, less obvious application of statistical NLP, which corresponds with avoidance of ambiguity: if C is by far the most plausible interpretation of F then there is no problem, since it is unlikely that F is

[7] See, e.g., Jurafsky 1996 and Jurafsky (to appear, section 4.2), where some reported mismatches between corpus frequencies and frequencies derived from psycholinguistic experiments are explained away.

[8] Cf., Blutner (1998), van Rooy (to appear), where the probability of a Content C is understood in terms of the likelihood of C *being true*. In light of the need for backing off, it will also be understood that the notations $p(F|C)$ and $p(C|F)$ are not intended in their strict sense, because they involve an abstraction over some suitable *class* of Contents and Forms, not C and F themselves.

misunderstood; if it is *not* by far the most plausible interpretation of F, then there is a problem. The notion of being 'by far the most plausible interpretation' (e.g., Abney 1996) can be formalised in different ways. For simplicity, we will take it to mean that C beats its nearest competitor by some sufficient margin. (If C beats its main competitor C' by only a slim margin then there is a fair chance that F is misunderstood as C'; the problem increases as the advantage of C decreases.) This idea can be formalised by saying that F is *viciously ambiguous* (henceforth, 'VA') with respect to Content C if there is another Content C' such that $t \cdot p(C|F) < p(C'|F)$, for some suitable threshold t such that $0 < t < 1$.[9] It will be useful to distinguish between being Viciously Ambiguous with respect to a given Content, and being Viciously Ambiguous *simpliciter*:

> Let $\mathcal{E}(C, F)$. Then, firstly, F is *viciously ambiguous* with respect to C iff there exists another Content C' such that $\mathcal{E}(C', F)$, where $t \cdot p(C|F) < p(C'|F)$. Secondly, F is *viciously ambiguous* iff F is VA with respect to all C' such that $\mathcal{E}(C', F)$.

Note that F may be VA with respect to one of its less probable interpretations without being (simply) VA. These definition schemas assume, of course, that Contents such as C and C' are sufficiently different that the choice matters; henceforth, we will assume that different Contents are never logically equivalent. The notion of vicious ambiguity might be used *directly* by, whereever this is possible, filtering out sentences that are VA with respect to the intended interpretation from the output of a generator, leading to an empirically-based version of the 'Generate and Revise' algorithm (section 3). Here, we will explore a more drastically bidirectional approach that takes the ordering of Forms as well as the ordering of Contents into account.

Before doing this, a possible objection might be mentioned: it might be thought that a corpus-based setting obliterates the need for bidirectionality, because $p(F|C)$ captures everything about F that is worth capturing when generating from C, including F's degree of ambiguity. This, however, would not only fail to explain the separate role of ambiguity, it is also not feasible in practice, since corpus-based probabilities will only capture broad regularities that bypass disambiguating factors. Imagine a corpus without ambiguity. In such a corpus, unambiguous conjunctions like *'old men and children'* may occur any number of times, and this could easily be taken as (false) legitimation for generating NPs like *'old men and women'*, which are ambiguous. This will not happen in a bidirectional approach, provided it is

[9] Note the difference between Vicious Ambiguity and Van Rooy's 'true ambiguity': a sentence is truly ambiguous if it has at least two interpretations that are optimally *relevant* (i.e., have optimal utility in a decision-theoretic sense) (van Rooy, to appear).

based on a grammar that attributes fewer interpretations to *'old men and children'* than to *'old men and women'.*

6 Towards a computational version of bidirectional superoptimality

To turn bidirectional superoptimality into a computational principle, our first move is to give computational meaning to OT-style expressions of the form $(\alpha, \beta) \prec (\gamma, \delta)$, where α and γ are Contents, and β and δ linguistic Forms. As anticipated in the previous section, we will interpret this by looking at probabilities of the forms $p(F|C)$ and $p(C|F)$. To the extent that the probabilities are derived from a corpus they are, of course, dependent on the linguistic style and *genre* of the corpus. Note that we only have to compare probabilities of pairs that share one of their arguments (that is, where $\alpha = \gamma$ or $\beta = \delta$), since these are the ones that compete with each other. In the new, probabilistic setting, superoptimality can be defined in many different ways (see also section 11.2). For concreteness, we will focus on a version of superoptimality that relies heavily on the notion of Vicious Ambiguity, and explore the consequences of this move, which follows on naturally from the discussion in the previous section:

> **Probabilistic superoptimality.** Let $\mathcal{E}(C, F)$. Then
> (C, F) is (probabilistically) superoptimal iff
> **(a)** there is no F' such that $p(F|C) < p(F'|C)$, and
> **(b)** F is not VA with respect to C.

Note the asymmetry between the two clauses: clause **(a)** stipulates that a Form F' can only cause F to be non-optimal if $p(F|C) < p(F'|C)$; clause **(b)**, on the other hand, implies that a Content, C' can cause F to be non-optimal even if C' is less probable than C, unless the difference in probability is substantial (i.e., unless $p(C'|F) \leq t \cdot p(C|F)$, see section 5). The reason for this asymmetry is that ambiguity can cause misunderstandings even if the intended interpretation is slightly more probable than all others. For example, suppose a corpus attaches the intended interpretation C_1 slightly more often to a given Form F than an alternative C_2 (e.g., $p(C_1|F) = 0.3$ and $p(C_2|F) = 0.29$). There is no reason to assume that every speaker of the language would always prefer C_1 over C_2 as an interpretation of F. The situation is different for probability of *Forms*: the most probable Form is always a good choice (until ambiguity is taken into account).

The notion of probabilistic superoptimality will prove to be useful, but additional ideas are needed. In particular, it would not be realistic to require that every Content C be expressed through a Form F such that (C, F) is probabilistically superoptimal. The problem, which we will dub the problem of *spurious ineffability*, is that some Contents may not be superoptimally expressible even though they *are* expressible. Suppose, for example, that we have a tiny language containing only two Contents, C_1 and C_2, and one Form F. Suppose $\mathcal{E}(C_1, F)$ and $\mathcal{E}(C_2, F)$, while $(C_1, F) \prec (C_2, F)$. In this case, there simply does not exist a superoptimal way of expressing C_1:

```
C2 ----- F
       /
      /
V    /
    /
   /
C1
```

OT theorists would argue that such situations correctly reflect the actually occurring phenomenon of *ineffability*: that some Contents may not be expressible in a given language. An example is discussed in Zeevat (2000, p.244 and p.257), where it is argued that the English sentence *'Who ate what?'* cannot be translated into a proper Italian sentence, as a result of a syntactic constraint that forces all Italian WH words to be fronted. Our situation is different, however, since our *selection* perspective (section 2) implies that we (unlike standard OT) are assuming $\mathcal{E}(C_1, F)$ to mean that F *is* a legitimate way of expressing E, the only question being whether it is the best way. Interpreted in terms of Eisner (2002)'s comparison with children and cloths (section 4), *not* dressing a child is not an option, unless there are no cloths that fit the child; similarly, *not* expressing a given Content is only an option if the grammar itself prevents it.

In the next section, we will explain how one can prevent spurious ineffability while still punishing ambiguity and accounting for blocking. The basic intuition will be that situations like the one described above should not prevent F from expressing C_1, since the language contains no *better* way of expressing C_1. This idea may be implemented in different ways, depending on the interpretation of 'better' (below: 'preferred over').

7 Strong superoptimality (Generation Perspective)

The key to making probabilistic superoptimality suitable for present purposes lies in distinguishing between superoptimal *expressions* and superoptimal *interpretations* without giving up bidirectionality. Let us start with a generation-oriented perspective, leaving the interpretation-oriented perspective till the next section. We define a novel notion of 'superoptimal expression', alongside the earlier-defined notion of probabilistic superoptimality (section 6).

> **Superoptimal expression.** Let C be a Content. Then F is a superoptimal expression of $C \Leftrightarrow_{def} \mathcal{E}(C, F)$ and there does *not* exists a Form G such that (C, G) is *preferred* over (C, F).

The notion of being *preferred* over can be defined in different ways. Bidirectionality suggests that a pair (C, G) can be preferred over a pair (C, F) for two kinds of reasons: one has to do with the ordering between Forms, the other with the ordering between Contents. One way to balance the two is as follows: (C, G) is *preferred* over (C, F) if it is preferred for one of the two types of reason, while the other type does not indicate the opposite preference.[10] The *first* type of possible reason for preferring (C, G) over (C, F) is that $p(F|C) < p(G|C)$, while the *second* is that F is more ambiguous than G with respect to Content C, in the following sense:

> **More ambiguous.** F is more ambiguous than G with respect to C $\Leftrightarrow_{def}$ F is VA with respect to C while G is not.

The notion 'preferred over' can now be defined succinctly:

> **Preferred Form.** Let $\mathcal{E}(C, F)$ and $\mathcal{E}(C, G)$.
> (C, G) is *preferred* over $(C, F) \Leftrightarrow_{def}$ either (1) or (2) or both:
> **(1)** $p(F|C) < p(G|C)$, and it is not the case that G is more ambiguous than F with respect to C, or
> **(2)** F is more ambiguous than G with respect to C, and it is not the case that $p(F|C) > p(G|C)$.

Clause (1) represents the case where (C, G) is preferred for a reason of the first type ($p(F|C) < p(G|C)$); clause (2) involves a reason of the second type (F is more ambiguous than G with respect to C). Note that both clauses consist of two conjuncts, the second of which guarantees, among other things, that the relation 'preferred over' is irreflexive and asymmetric.

[10] Since only two reasons are taken into account, this is consistent with various well known decision-theoretic strategies, including: *Plurality* (i.e., (C, G) is *preferred* over (C, F) if more reasons favour (C, G) over (C, F) than the other way round) and *Pareto optimality* (i.e., (C, G) is *preferred* over (C, F) if at least one reason favours (C, G) over (C, F), while no reason favours (C, F) over (C, G)).

For illustration, compare the graph in the previous section with a slightly larger one. In the earlier graph, F was a probabilistically superoptimal expression of C_1 since there was no pair (C_1, G) that is preferred over the pair (C_1, F). This was true even though the pair (C_1, F) was not probabilistically superoptimal (in the sense of section 6). In the expanded graph, however, (C_1, F) is 'blocked' by (C_1, G) because (C_1, G) is preferred over (C_1, F).

```
C2 ----- F
       /
      /
V    /   =
    /
   /
C1 ----- G
```

A situation of this kind might occur, for example, with the input Content $C_1 = \exists y \forall x(Love(x, y))$, and the 'distractor' Content $C_2 = \forall x \exists y(Love(x, y))$, if C_1 is expressed equally well by $F =$ *'Every man loves a woman'* and by $G =$ *'There is a woman who is loved by every man'*, (making the usual assumption that F is ambiguous, whereas G can only mean C_1). A variant of this example arises if we assume that F is more fluent than G, in which case $p(F|C_1) > p(G|C_1)$, causing the vicious ambiguity of F to be offset by the awkwardness of G, making both F and G superoptimal expressions of C_1.

A quasi-procedural perspective on the different ways in which a given Content may be expressed will help us prove our expressibility theorem (Theorem 1, below). Imagine a sequence $\mathcal{F}$ of Forms $F_1, F_2, F_3, ...$, starting with $F_1 = F$, where $\mathcal{E}(C, F_i)$ for all $i > 0$, and where F_{i+1} constitutes an improvement over F_i in the sense that (C, F_{i+1}) is preferred over (C, F_i). $\mathcal{F}$ will be called an *improvement sequence.* An improvement *step* from F_i to F_{i+1} can be of three kinds: it can arise from clause (1) of the definition of 'preferred' (i.e., $p(F_{i+1}|C) > p(F_i|C)$, while F_{i+1} is not more ambiguous than F_i), from clause (2) (i.e., F_{i+1} is less ambiguous than F_i, while $p(F_{i+1}|C) \geq p(F_i|C)$), or from both clauses at the same time (i.e., $p(F_{i+1}|C) > p(F_i|C)$ and F_{i+1} is less ambiguous than F_i). An improvement step of the first kind will be called an *abbreviation* step (because a more probable Form will often be shorter); a step of the second kind will be called a *disambiguation* step; a step of the third kind is an abbreviation step as well as a disambiguation step.

For illustration, let us look at a schematic rendering of a possible situation involving a grammar relating the forms $F_1, .., F_{10}$ to the same Content, say C. The letter a stands for an abbreviation step, d stands for a disambiguation step, ad stands for a step that abbreviates and disambiguates. (Thus, F_7 disambiguates F_1, F_8 abbreviates F_7, etc.)

```
      F2  a  F3                      F6
   a              ad              a
F1                     F4  a  F5
   d                 a            a
      F7 a F8 a F9                   F10
```

This diagram condenses four different improvement sequences originating in F_1, namely $\langle F_1, F_2, F_3, F_4, F_5, F_6\rangle$, $\langle F_1, F_2, F_3, F_4, F_5, F_{10}\rangle$, $\langle F_1, F_7, F_8, F_9, F_4, F_5, F_6\rangle$, and $\langle F_1, F_7, F_8, F_9, F_4, F_5, F_{10}\rangle$. Crucially, none of these sequences contains multiple occurrences of the same Form, and each sequence is finite. In the situation depicted, both F_6 and F_{10} are superoptimal expressions of C, and no other Form is. (In the case at hand, this can only be because $p(F_6|C) = p(F_{10}|C)$.) An improvement sequence may sometimes result in an end point that is still VA, but this can only happen if no disambiguation step is applied. For example, the step from F_1 to F_7 implies that F_1 is VA while F_7 is not; consequently, none of $F_8, F_9, F_4, F_5, F_6, F_{10}$ can be VA.

We are now ready to prove a first theorem, making use of the idea of an improvement sequence. A relation Q will be called well founded if there does not exist an infinite chain $x_1, x_2, x_3, ...$ such that $x_i Q x_{i+1}$ for all natural numbers i (cf., Landman 1991, chapter 2; also Jäger 2001).

> *Theorem 1:* For any Content C and Form F, if $\mathcal{E}(C,F)$ then there exists a *superoptimal* expression of C. *Proof*: Let the notation $F\mathcal{R}F'$ mean that $\mathcal{E}(C,F)$ and $\mathcal{E}(C,F')$ and (C,F') is *preferred* over (C,F). The theorem follows directly from the following Lemma.
>
> *Lemma:* $\mathcal{R}$ is well founded. This follows from the following two propositions.
>
>> *Proposition 1.* (Disambiguation step.) An improvement sequence can contain at most one disambiguation step. *Proof:* After a disambiguation step, the resulting F_{i+1} is not VA. As a result, no later Form in the sequence (i.e., no F_j for $j > i + 1$) can be VA.

> *Proposition 2.* (Abbreviation steps.) The relation $\prec$ is well founded in the sense that, for every Content C, there does not exist an infinite series $F_1, F_2, ...$ such that $p(F_{i+1}|C) > p(F_i|C)$ for all i. *Proof:* This follows directly from the probabilistic setting, where the sum of the probabilities $p(F|C)$, for all F such that $\mathcal{E}(C, F)$, equals 1.[11]

These two Propositions imply that $\mathcal{R}$ must be well founded. To see this, suppose there existed an infinite improvement sequence $F_1, F_2, F_3, ...$ such that $F_i \mathcal{R} F'_{i+1}$ for every i. If this sequence does not contain any disambiguation steps, its existence contradicts Proposition 2. But if it does contain disambiguation steps then Proposition 1 tells us there can be only one such step; suppose this is the step from F_i to F_{i+1}, then the sequence $F_{i+1}, F_{i+2}, F_{i+3}, ..$ is still infinite, without any disambiguation steps, contradicting Proposition 2 once again. QED

The problem of spurious ineffability formed the main remaining obstacle to applying *probabilistic superoptimality* to NLG, and this problem has now been tackled.

8 Strong superoptimality (Interpretation Perspective)

Although this paper focuses mainly on generation, our approach has a mirror image in interpretation, as will be explained in this section. Our reasoning will be analogous to that in the previous section. The starting point is the observation (mirroring observations in section 6) that, for some interpretable Forms F_1, there may not exist an interpretation C such that (C, F_1) is probabilistically superoptimal:

[11] In the usual OT setting, the well-foundedness of the ordering of $\{F : \mathcal{E}(C, F)\}$ follows from the way in which the ordering is derived from a set of constraints (Jäger 2002).

```
C  ----- F1
   \
    \
     \   ^
      \
       \
         F2
```

Clearly, probabilistic superoptimality of an entire pair (C, F) is too strong a requirement. Therefore, we take the input F as given and define the notion of a superoptimal *interpretation* of a given Form, analogous to the notion of a superoptimal *expression* of a given Content.[12]

> **Superoptimal interpretation.** Let F be a Form. Then C is a superoptimal interpretation of F $\Leftrightarrow_{def}$ $\mathcal{E}(C, F)$ and there does not exist a Content D such that (D, F) is *preferred over* (C, F).

This definition relies on another definition, spelling out what it means to say that (D, F) is *preferred* over (C, F). Analogous to section 7, we will distinguish two types of reasons why (D, F) may be preferred over (C, F). The *first* is that $p(C|F) < p(D|F)$, while the *second* is that C could have been expressed better than via F, while D could not. This is defined as follows:

> Let $\mathcal{E}(C, F)$. Then we will say that C 'could have been expressed better' $\Leftrightarrow_{def}$ $\exists F' : \mathcal{E}(C, F')$ & $p(F|C) < p(F'|C)$.

Giving equal weight to the two different types of reasons as was done in section 7, we arrive at the following definition of what it means for (D, F) to be preferred over (C, F):

> **Preferred Content.** Let $\mathcal{E}(C, F)$ and $\mathcal{E}(D, F)$.
> (D, F) is *preferred* over (C, F) $\Leftrightarrow_{def}$ either (1) or (2) or both:
> **(1)** $p(C|F) < p(D|F)$, and it is not the case that
> (D could have been expressed better while C could not)
> **(2)** (C could have been expressed better while D could not)
> and it is not the case that $p(C|F) > p(D|F)$.

Clause (1) represents the case where (C, G) is preferred for a reason of the first type (i.e., $p(C|F) < p(D|F)$); clause (2) involves a reason of the second type. Analogous to section 7, the new relation of 'preferred over' is irreflexive and asymmetrical. The rationale for clause (1) can be seen in the following

[12] Because, this time around, *interpretation* is the starting point, there will not be a role for vicious ambiguity (unlike the *generative* mechanism of the previous section), until more complex interactions between interpretation and generation are considered, at the end of section 11.2.

example, where both D and C are superoptimal interpretations of F, since each of the two interpretations has a drawback: (D, F) is not preferred over (C, F) (because D could have been expressed better, while this is not true for C) nor is (C, F) preferred over (D, F) (because $p(D|F) > p(C|F)$).

```
D  ----- F'
    \
     \
v     \   v
       \
        \
C  ----- F
```

Given this understanding of what it means to be a superoptimal interpretation, anything that can be interpreted can be interpreted superoptimally:

> *Theorem 2:* For any Content C and Form F, if $\mathcal{E}(C, F)$ then there exists a superoptimal interpretation of F.
>
> *Proof:* Analogous to Theorem 1. Let $C\mathcal{Q}C'$ be the relation holding between Contents C and C' if $\mathcal{E}(C, F)$ and $\mathcal{E}(C', F)$ and (C', F) is *preferred over* (C, F). Direct analogons of Propositions 1 and 2 are easily proven (based on the two types of improvement steps that arise from the definition of **Preferred Content**). These Propositions imply that there cannot exist an infinite chain $x_1, x_2, x_3, ...$ such that $x_i\mathcal{Q}x_{i+1}$ for all i. Theorem 2 follows directly from this fact. QED

9 A worked example

To see how superoptimal expression and interpretation dovetail, let us look at one example, using some fictional probablities for concreteness. Suppose the relevant Contents and Forms are as follows while $\mathcal{E}(C_1, F_1)$, $\mathcal{E}(C_1, F_2)$, $\mathcal{E}(C_2, F_2)$, and $\mathcal{E}(C_2, F_3)$. We will use fully specified interpretations and write them in ordinary predicate logic.

Contents:

$C_1 = \forall x((Man(x) \vee Woman(x)) \rightarrow (Old(x) \rightarrow Ben(x)))$
$C_2 = \forall x(Woman(x) \rightarrow Ben(x))$ & $\forall x(Man(x) \rightarrow (Old(x) \rightarrow Ben(x)))$

Forms:

$F_1 =$ *'Old men and old women are entitled to benefits'* (nonambiguous)
$F_2 =$ *'Old men and women are entitled to benefits'* (ambiguous)
$F_3 =$ *'Women and old men are entitled to benefits'* (nonambiguous)

Probabilities:

$p(F_1|C_1) < p(F_2|C_1)$
$p(F_2|C_2) = p(F_3|C_2)$
$p(C_1|F_2) = p(C_2|F_2)$

```
C1 ----- F1 (unambiguous)
   \
    \
=    \   ^
      \
       \
C2 ----- F2 (ambiguous)
   \
    \
     \   =
      \
       \
         F3 (unambiguous)
```

First, let us take a generative perspective. There are two possible inputs to generation:

> *Input* C_1. Both F_1 and F_2 are superoptimal expressions of C_1 since $p(F_1|C_1) < p(F_2|C_1)$ while, on the other hand, F_2 is VA with respect to C_1. [N.B. If the example is modified by assuming that $p(F_1|C_1) = p(F_2|C_1)$ then only F_1 is a superoptimal expression of C_1, since F_2 (not F_1) is VA with respect to C_1.]
> *Input* C_2. Obviously, F_3 is a superoptimal expression of C_2. F_2 is not a superoptimal expression of C_2, since F_2 is VA with respect to C_2, whereas there is an alternative, F_3, which is not.

Now, an interpretive perspective. There are three possible inputs.

> *Input* F_1 *or* F_3. These Forms have only one possible interpretation, which must therefore be a superoptimal interpretation.
> *Input* F_2. C_1 and C_2 are both superoptimal interpretations of F_2, since neither (C_1, F_2) nor (C_2, F_2) is preferred over the other. [N.B. If the example is modified by assuming that $p(F_2|C_1) < p(F_1|C_1)$ then only C_2 is a superoptimal interpretation of F_2, since C_1 could have been expressed better, but C_2 could not.]

The example shows that the generator assumes the interpreter to be rational (i.e., avoiding dispreferred interpretations), while the interpreter assumes the generator to be rational (i.e., avoiding dispreferred Forms). This is, of course, entirely in the Gricean spirit of bidirectional OT. Note that interpretation and generation are not each other's inverses, since C_2 is a superoptimal interpretation of F_2, but F_2 is not a superoptimal expression of C_2 (cf., Smolensky 1996, Zeevat 2000).

10 Algorithms and complexity

The computational implications of the approach sketched in the previous sections may be worth exploring briefly. Previous research on other variants of OT has shown that it is not always necessary to compute all the different ways in which a given Content may be expressed, since grammar and constraints may be compiled into one Finite State Transducer (Frank and Satta 1998 for the unidirectional case; Jäger 2002 for the bidirectional case), and something similar might hold for the ideas presented here. For now, however, let us assume a straightforward implementation along the following lines: In generation, the first step is to compute the set Forms_C (see section 4), consisting of all the Forms F that can express the input Content C; for each of these pairs (C, F), the probability $p(F|C)$ is calculated. The second step is to calculate, for each element F of Forms_C, the set Contents_F, consisting of all the Contents expressible by F. Choosing a suitable expression of the original input C is now a matter of eliminating all those F such that F is not a superoptimal expression of C and making an arbitrary choice from the remaining alternatives if there is more than one.

Superoptimal Generation from input C

1. Compute $Forms_C$, keeping track of probabilities $p(F|C)$
2. For every $F \in Forms_C$, compute $Contents_F$,
   keeping track of probabilities $p(C|F)$
3. Choose $G \in Forms_C$ such that G is a superoptimal expression of C.
   The relevant set of Forms and Contents
   (with respect to which superoptimality is calculated)
   is $Forms_C \;\bigcup\; \cup_{F \in Forms_C} Contents_F$.

Superoptimal Interpretation is the mirror image of this procedure.[13]

Assuming an implementation along these lines, the number of calculations

[13] Note that Zeevat 2000 (sections 5 and 6) although superficially similar, appears to make very different predictions. In Zeevat's approach to the generation from a Content C, for example, a Form F can never be selected if another Form F' violates fewer (generation) constraints than F, nor if another Content C' violates fewer (interpretation) constraints than C, which appears to make successful generation something of a miracle. The situation for interpretation is essentially the same. Our own account, with its emphasis on superoptimal *expression* and superoptimal *interpretation*, makes 'successful' generation/interpretation commonplace, as is proven by Theorems 1 and 2.

performed depends on the number of Contents associated with any given Form, and on the number of Forms associated with any given Content. The worst-case complexity of the just-presented algorithm may be assessed in the following way. Let f be the maximum number of Forms that correspond with any given Content (i.e., the maximum number of Forms F such that $\mathcal{E}(C, F)$, for any Content C). Likewise, let c be the maximum number of Contents that correspond with any given Form (i.e., the maximum number of Contents C such that $\mathcal{E}(C, F)$, for any Form F). Then, in order to check whether F is a superoptimal expression of C, at most $f.c$ combinations have to be inspected (going from a Content to all its corresponding Forms, then back to all their corresponding Contents) leading to a complexity of $O(fc)$. Checking whether C is a superoptimal interpretation of F has the same complexity, making complexity polynomial in both cases.[14]

These results do not tell us how much time it takes to generate a Form given a certain Content (and the other way round for interpretation), and this time could be considerable. Also, f and c could prove to be huge. Various strategies might be explored to cope with these issues. (For example, one could look for shortcuts, e.g., letting step (3) of the algorithm disregard all but the 10 most probable C in $Forms_F$.) This paper is not the place for detailed solutions to these issues, which have haunted other areas of OT as well.

11 Conclusion

We will sum up the essentials of the approach outlined in this paper, after which some alternatives to the specific proposal in sections 7 and 8 will be sketched. The paper concludes with some remarks about future work.

11.1 Summing up

In this paper, a connection has been made between bidirectional Optimality Theory (OT) and Natural Language Processing. More specifically, we have explored how OT can be connected with current ideas on the avoidance of ambiguity in NLG systems. We believe this to be of interest because, so far, bidirectional OT syntax and semantics have been a predominantly

[14] If a third reason for preferring (D, C) over (C, F) (see section 11) is adopted, then complexity of interpretation goes up to $O(fc^2)$, since these interactions would go one level deeper than those proposed in section 8.

theoretical affair whereas the avoidance of ambiguity in NLG, when it is addressed at all[15], has not taken the plausibility of the relevant interpretations into account. The connection between these two areas of semantic research suggests adaptations of OT that allow Natural Language *Generation* systems to balance the different requirements on generated text in a principled way, by taking interpretation into account; similarly, it offers a perspective on Natural Language *Understanding* in which the interpreter takes into account which other sentences might have been uttered. Together, these developments suggest a possible future in which computational generation and interpretation are inextricably linked.

We have used only a few of the ideas commonly associated with OT. For example, we have said little about constraints, advocating corpus-based methods for inducing orderings of Contents and Forms instead (section 5). Also, we have refrained from making any claims about language universals, as is often done in OT; quite the contrary, our corpus-based based approach would tend to associate different orderings of Contents and Forms not only with different languages, but even with different genres. The key element of OT that does survive, however, is the idea of bidirectionality, as proposed by Blutner (2000) and Jäger (2002).

11.2 Variants of our approach

The ideas outlined in the previous paragraphs can be made precise in different ways and it is unclear, at this stage, which of them is most fruitful. The proposals of the previous sections should therefore be regarded as merely illustrative. Let us briefly sketch some variants of these proposals, and highlight some of their advantages and disadvantages.[16]

An important limitation of the approach of sections 7 and 8 stems from the fact that all cases of VA are lumped together, as if they were all equally serious; similarly, all Form that are *not* VA are also lumped together. This dychotomy between VA and not-VA can sometimes distort reality. Some informal new terminology might be useful. Let F^* be a Form whose conditional likelihood given C is maximal in the sense that no other element of $\{F : \mathcal{E}(C, F)\}$ has greater likelihood; similarly, let C^* be an element of

[15] Witness Reiter and Dale (2000), where avoidance of ambiguity is only discussed in connection with the generation of referring expressions (although revision architectures are briefly mentioned in sections 3.5.2 and 3.6).

[16] A number of these variants, and the considerations leading up to them, were suggested to me by one of the reviewers of Journal of Semantics.

$\{C : \mathcal{E}(C, F)\}$ whose conditional likelihood given F is maximal. Then one might say that a Form F is only *slightly* VA with respect to a content C if C is roughly equally likely as C^*, whereas F is *highly* VA if it far less likely. Furthermore, one might call F *almost* VA if its most likely interpretation C^* beats its nearest competitors by a margin just large enough to make F not VA. With these informal distinctions in place, let us look at the predictions that our account is making, bearing in mind that any given Form must be VA with respect all except (possibly) one of its interpretations.

Firstly, consider a situation in which F is *slightly* VA while G is *almost* VA. Then if $p(G|C) < p(F|C)$, both F and G will count as superoptimal expressions of a Content C, even if $p(G|C) << p(F|C)$, simply because F is VA while G is not. Arguably, this outcome is too generous towards the Form G, given that its large disadvantage regarding fluency/brevity is not offset by its marginal advantage regarding ambiguity. Conversely, suppose F and G are both VA with respect to C. Then even the slightest 'formal' advantage of G over F (i.e., when $p(G|C)$ slightly exceeds $p(F|C)$) will cause G to be preferred over F, even if F is only *slighly* VA while G is *highly* VA with respect to C. But in a situation like this, F should arguably be preferred over G:

```
C'
   \
    \
     \
=      F
     /
    /
   /
C /    ^
  \
   \
    \
^    \
^      G
     /
    /
   /
C''
```

In light of this disadvantage, let us explore some alternative approaches, which do not rely on the notion of VA. A neatly symmetrical approach results if superoptimality is applied to OT directly, without letting VA play a role. Superoptimality could either be defined along Blutnerian lines (i.e., using constraints) or probabilistically, to avoid reliance on constraints:

> **Probabilistic superoptimality (Variant (i)).** Let $\mathcal{E}(C, F)$. Then (C, F) is superoptimal iff
> **(a)** there is no F' such that $p(F|C) \ << \ p(F'|C)$, and
> **(b)** there is no C' such that $p(C|F) \ << \ p(C'|F)$.

Variant **(i)** can be made more specific by focussing on the above-defined favourite Form F^* and favourite Content C^*, using thresholds t_1 and t_2 (whose values, between 0 and 1, would have to be determined):

> **Probabilistic superoptimality (Variant (ii)).** Let $\mathcal{E}(C, F)$. Then (C, F) is superoptimal iff
> **(a)** $p(F|C) \ > \ t_1 \cdot p(F^*|C)$ and
> **(b)** $p(C|F) \ > \ t_2 \cdot p(C^*|F)$.

In words: a pair (C, F) is superoptimal if F is almost maximally likely given C, while C is almost maximally likely given F. An account along these lines may be appealingly simple, and it would avoid lumping together all cases of VA, but it has some drawbacks of its own, depending on whether superoptimality is applied directly, or via the notion of a superoptimal *expression*.

First, suppose superoptimality is applied *directly*. In this case, the problem of spurious ineffability (and the analogous problem for interpretation) returns, since not every Content C may be expressible through a Form F such that the pair (C, F) is superoptimal. (Examples showing this, exactly like those in sections 6 and 8, are easily constructed.) Two different responses to the problem of spurious ineffability are possible. The first is to let the generator *reject* contents that cannot be expressed via a superoptimal pair: where there does not exist an expression that is good enough, the aggregation module of the generator should propose a more suitable Content, for example by dividing C over two different utterances. In generation, this might be a viable approach, but it is very doubtful whether the analogous response regarding the *interpretation* of a problematic Form is realistic, since interpretation has to work with the input that it is given.

The second possibility is to use the above-defined notion of superoptimality *indirectly*, by invoking the notion of a superoptimal *expression*:

> **Superoptimal expression (Repeated from section 7).** Let C be a Content. Then F is a superoptimal expression of $C \Leftrightarrow_{def} \mathcal{E}(C, F)$ and there does *not* exists a Form G such that (C, G) is *preferred* over (C, F).

It would now be natural to let this definition rest on the following revised notion of being preferred:

> **Preferred Form (alternative).** Let Let $\mathcal{E}(C, F)$ and $\mathcal{E}(C, G)$.
> (C, G) is *preferred* over (C, F) $\Leftrightarrow_{def}$
> (C, G) is superoptimal and (C, F) is not.

In this way, spurious ineffability could be finessed. But there is another problem, since not all cases of VA are captured by the alternative approach: according to the new definitions, a pair (C, F) counts as superoptimal even if F has an interpretation D that is roughly equally likely as C. In the following situation, for example, let us assume that F and F' are roughly equally likely as formulations of the input C; similarly, C and D are roughly equally likely as interpretations of F:[17]

```
C  ----- F'
    \
     \
=     \   =
       \
        \
D  ----- F
```

According to section 7, F's being VA counts against it, causing F' to be the favoured formulation of C. The alternative account, suggested above, would only 'punish' F if it had an interpretation that was significantly *more* likely than C. Based on the discussion in section 5 the original account seems preferable in this respect. The battle between VA based and non-VA based approaches appears to end in a draw.

A different type of criticism that could be made of all the proposals discussed so far is that they simplify the interaction between interpretation and generation in an arbitrary way. Taking an interpretive point of view, for a change, let us consider the following situation, which might obtain in American English if *'you'* denotes either singular or plural, while *'you guys'* denotes the plural only.

[17] We take it as understood that, in this diagram, it is $p(C|F)$ and $p(D|F)$ (not $p(C|F')$ and $p(D|F)$) that are compared. If the lines of a diagram were to cross, diagrams of this kind would become ambiguous and some other type of notation would have to be used.

```
[plural referent]     C  ----- G    (‘You guys’)
                          \
                           \
                      =     \   =
                             \
                              \
[singular referent]   D  ----- F    (‘You’)
```

The theory of section 8 does not allow us to choose between C and D; but one might argue that C should be blocked, because this Form could have been expressed without vicious ambiguity (which would have been preferable – so the argument would go), while this is not true for D itself. This effect could be achieved by adding a third reason why (D, F) can be preferred over (C, F) to the mechanism of section 8, namely that

$\forall H(\mathcal{E}(D, G) \Rightarrow (H \text{ is VA with respect to } D))$ whereas
$\neg\forall H(\mathcal{E}(C, H) \Rightarrow (H \text{ is VA with respect to } C))$.

Whether more complex interactions of this kind lead to more accurate predictions concerning the preferred interpretation of a Form (or, analogously, concerning the preferred formulation of a Content), we do not know.

11.3 Future work

Having discussed a number of ways in which ideas from OT can inform NLP (and more specifically NLG), we have observed that – rather characteristically in connection with OT – each of the alternatives has its own advantages and disadvantages; which variant works best is something that only further research can determine. Any of the mechanisms discussed could be superimposed on an actual NLP system, as long as the system allows different expressions of the same Content and different interpretations of the same Form, provided empirically supported probabilities $p(F|C)$ and $p(C|F)$ are available, for which – as was argued in section 6 – we believe there is now a real prospect. A language generation system that used the methods proposed here would have to be evaluated to determine whether they balance naturalness and clarity in a proper way, and this would be likely to lead to modifications of the method. An important open question is how to let linguistic context play a role, since this is an aspect that we have allowed ourselves to ignore (cf., footnote 1).

12 Acknowledgment

This paper grew out of two recent experiences. One is my exposure to bidirectional OT at ESSLLI-2002 in Trento, as represented by Paul Dekker and Robert van Rooy. The other is a series of discussions on wide-coverage NLG over the last year or so in Brighton, involving Anja Belz, John Carroll, Roger Evans, and David Weir. The idea of avoiding only those ambiguities that are likely to be misunderstood given the predictions of a statistical parser is due to my colleagues at the University of Sussex. Thanks are also due to David Ahn, Ted Briscoe, John Carroll, Emiel Krahmer, the audience of the Fifth International Workshop on Computational Semantics (IWCS-5) and, last but not least, the anonymous reviewers of Journal of Semantics, whose advice I have followed in many cases. Support from the UK's Engineering and Physical Sciences Research Board (EPSRC) under grant number GR/S24480 (COGENT) is gratefully acknowledged.

Kees van Deemter
Information Technology Research Institute (ITRI)
University of Brighton, Lewes Road, Watts Building
Brighton BN2 4GJ, United Kingdom

Email: Kees.van.Deemter@itri.brighton.ac.uk
Web: http://www.itri.brighton.ac.uk/∼Kees.van.Deemter/
Tel: +44 1273 642910
Fax: +44 1273 642908

13 References

Abney (1996) S.Abney. 'Statistical Methods and Linguistics'. In J.Klavans and Ph.Resnik (eds.) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* The MIT Press, Cambridge, MA. 1996.

Beaver (in press). 'The Optimization of Discourse Anaphora'. To appear in *Linguistics and Philosophy.*

Blutner (1998) 'Lexical Pragmatics'. *Journal of Semantics* 15(2).

Blutner (2000) 'Some aspects of optimality in Natural Language Interpretation'. *Journal of Semantics* 17(3).

Boersma (1997) P.Boersma. 'How we learn variation, optionality, and probability'. In Proceedings of the Institute of Phonetic Sciences, Amsterdam (IFA) 21: 43-58. Also in Rutgers Optimality Archive 221.

Bresnan (in press). J.Bresnan. 'Optimal Syntax'. In J.Dekkers, F.van der Leeuw and J.van de Weijer, *Optimality Theory: Phonology, Syntax and Acquisition.* To appear with Oxford University Press.

Briscoe, E. and J. Carroll (2002) 'Robust accurate statistical annotation of general text'. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Gran Canaria. p.1499-1504.

Burzio (1989). 'On the non-existence of disjoint reference principles'. *Rivista di Grammatica Generativa* 14, 3-27.

Butt et al. (1999). Miriam Butt, Tracy Holloway King, Maria-Eugenia Nino, and Frédérique Segond. 'A Grammar Writer's Cookbook'. CSLI Publications. Stanford, Ca.

Callaway (2003). Ch.Callaway. 'Multilingual revision'. To appear in Procs. of 9th European Workshop on Natural Language Generation. Budapest, April 2003.

Carroll et al. (2002) J.Carrol, R.Evans, K.van Deemter, D.Weir, and A.Belz. 'COGENT: COntrolled GENeration of Text'. Internal document, University of Brighton and University of Sussex.

Charniak (2000) E.Charniak. 'A maximum entropy inspired parser'. Procs. of first Conf. of the North American Chapter of the Association for Computational Linguistics, Seattle, WA, pp.132-139.

Chelba and Jelinek (2000) 'Structured language modeling'. Computer Speech and Language **14** (4).

Chomsky (1981) N.Chomsky. *Lectures on Government and Binding.* Foris, Dordrecht.

Copestake and Flickinger (2000) 'An open-source grammar development environment and broad-coverage English grammar using HPSG.' Procs. LREC.

Copestake et al. (1999) A.Copestake, D.Flickinger, and I.Sag. 'Minimal Recursion Semantics'. Technical note, CSLI, Stanford.

Dale and Reiter (1995) 'Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions'. *Cognitive Science* **18**, 233-263.

Dekker and Van Rooy (2000) P.Dekker and R.van Rooy. 'Bi-directional Optimality Theory: An Application of Game Theory'. *Journal of Semantics* 17 (3).

Eisner (2002) J.Eisner. 'Comprehension and compilation in Optimality Theory'. In Procs. of 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL-2002). Philadelphia.

Frank et al. (1998) A.Frank, T.Holloway King, Jonas Kuhn, and John Maxwell. 'Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars'. In M.Butt and T.Holloway King (Eds.), Procs. of the LFG98 Conference, Univ. of Queensland, Brisbane.

Frank and Satta (1998) R.Frank and G.Satta. 'Optimality Theory and the Generative Complexity of Constraint Violability'. *Computational Linguistics* **24**, p.307-315.

Harrison (1988) Ph.Harrison. *A new algorithm for parsing Generalized Phrase Structure Grammar.* Ph.D. Dissertation, University of Washington.

Hendriks and de Hoop (2001) 'Optimality Theoretic Semantics'. Linguistics and Philosophy 24.1, pp. 1-32

Inui et al. (1992) K.Inui, T.Tokunaga, H.Tanak. 'Text revision: a model and its implementation'. Proceedings of the 6th Int. workshop on Natural Language Generation.

Jäger (2002) G. Jäger. 'Some notes on the formal properties of bidirectional optimality theory'. *Journal of Logic, Language and Information* 11 (4).

Johnson et al. (1999) M.Johnson, S.Geman, S.Canon, Z.Chi, and S.Riezler. 'Estimators for stochastic 'unification-based' grammars'. In procs. of ACL-1997, pp.535-541. College Park, Maryland.

Jurafsky, Daniel. (1996) 'A Probabilistic Model of Lexical and Syntactic Access and Disambiguation.' Cognitive Science 20 137-194.

Jurafsky, Dan. (to appear) 'Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production'. To appear in Rens Bod, Jennifer Hay, and Stefanie Jannedy, (Eds)., *Probabilistic Linguistics.*

Kager, René (1999) *Optimality Theory.* Cambridge. Cambridge University Press.

Kingsbury et al. (2002) Paul Kingsbury, Martha Palmer, and Mitch Marcus. 'Adding Semantic Annotation to the Penn TreeBank'. In Proceedings of the Human Language Technology Conference, San Diego, California.

Knight and Langkilde (1998) K.Knight and I.Langkilde. 'Generation that exploits corpus-based statistical knowledge'. Procs. COLING-ACL, pp.704-710.

Kuhn (2003) Jonas Kuhn. *Optimality-Theoretic Syntax: A Declarative Approach: A Declarative Approach.* CSLI Publications, Studies in Constraint Based Lexicalism.

Landman (1991) F.Landman. *Structures for Semantics.* Kluwer Academic Publishers, Studies in Linguistics and Philosophy no.45.

Levelt (1989) W.J.M.Levelt. *Speaking.* The MIT Press, Cambridge Mass.

Meteer and McDonald (1986) A Model of Revision in Natural Language Generation. Proc. ACL 1986, p.90-96.

Neumann and van Noord (1992) 'Self-monitoring with reversible grammars'. In Procs. of 14th COLING conference.

Neumann (1994) G. Neumann. *A Uniform Computational Model for Natural Language Parsing and Generation.* PhD thesis, University of the Saarland, Saarbrcken.

Neumann (1998) G. Neumann: 'Interleaving Natural Language Parsing and Generation Through Uniform Processing.' Artificial Intelligence 99, (1998) pp. 121-163.

Power et al. (2003) Power, R. J. D., Scott, D. R., and Bouayad-Agha, N. (in press). 'Document Structure'. *Computational Linguistics* **29** (2).

Prince and Smolensky (1991) 'Notes on connectionism and harmony theory in linguistics'. Technical Report, Dept. of Computer Science, Univ. of Colorado, Boulder.

Prince and Smolensky (1997) Prince, Alan and Paul Smolensky. 'Optimality: From Neural Networks to Universal Grammar'. Science 275, 1604-1610.

Reiter and Dale (2000) *Building Natural Language Generation Systems.* Cambridge University Press.

Robin (1994) Automatic generation and revision of natural language summaries providing historical background. In procs. of 11 the Brazilian Symposium on Artificial Intelligence (SBIA-94).

Smolensky (1996) 'On the comprehension/production dilemma in child language. *Linguistic Inquiry* **27**.

van Deemter (2002) 'Generating Referring Expressions: Boolean Extensions of the Incremental Algorithm' *Computational Linguistics* **28** (1), pp.37-52. March 2002.

van Rooy (to appear) R. van Rooy. 'Relevance and Bidirectional OT'. To appear in R. Blutner and H. Zeevat (eds.), *Pragmatics in Optimality Theory*, Palgrave Macmillan.

Reiter and Dale (2000) *Building Natural Language Generation Systems.* Cambridge University Press, Cambridge.

Shieber (1993) S. Shieber. 'The Problem of Logical-Form Equivalence'. Squib in *Computational Linguistics* 19, 1.

Toutanova et al. (2002) K. Toutanova, C. D. Manning, S. M. Shieber, D. Flickinger, and S. Oepen. 2002. 'Parse Disambiguation for a Rich HPSG Grammar'. First Workshop on Treebanks and Linguistic Theories (TLT2002), pp. 253-263.

Yazdani (1987) 'Reviewing as a Component of the Text Generation Process'. In G.Kempen (ed.) *Natural Language Generation*, Chapter 13, p.183-190. Martinus Nijhoff.

Zeevat (2000) 'The asymmetry of Optimality Theoretic Syntax and Semantics' *Journal of Semantics* 17(3).

Zeevat (2002). 'Double Bias'. In Procs. of 7th Symposium on Logic and Language, Pécs, Hungary, Aug. 2002.