[From: Internat.conf. *Methodology & Techniques of Machine Translation*, British Computer Society, Cranfield, 13-15 February 1984]

How to Boldly Split Infinitives

Raphael Mankin

Solvfield Ltd.,
London, U.K.

INTRODUCTION

See the little phrases go,
Observe their funny antics;
The men who make them wriggle so
Are teachers of semantics.

In this paper I shall advocate that the normal procedure for computer analysis of language should be stood on its head. I shall also argue for a very simple syntax whose main function is to disambiguate items whose general meaning is already known, and to point out inherent ambiguity.

When one looks at standard linguistics there is much emphasis on syntax as the primary vehicle of analysis. Semantics only comes in to play after the syntactic structure of the utterance has been determined. In effect: syntax is what you can say, and semantics is what it means when you have said it. I shall argue for the inversion of this order; that the semantic analysis should come first. In terms of computer processing I am an advocate of the primacy of the lexicon, so that the semantics is what you are talking about, and syntax is what you are saying about it.

Having developed the semantics properly we shall find that very little syntactic processing is necessary. A large number of analyses will be ruled out by their being meaningless rather than by their being "ungrammmatical". I shall use the split infinitive and the general problem of the positioning of adverbs as the illustrative peg on which to hang the discussion.

COMPREHENSIBILITY AND COMPREHENSION

Twas brillig and the slithy toves
Did gyre and gimbal in the wabe.

Since a grammar has to deal with the comprehension of utterances it must, from the outset, consider the meanings of

words. Thus in processing, we must first ask what each word (or particle or phrase) means. The syntax has then to organize meaningful elements into larger meaningful constructs. In this way we can explain why some "syntactically aberrant" sentences are sensible, and other "syntactically correct" sentences are nonsense.

To boldly split infinitives where no man has boldly split before. ... 1

Colourless green ideas sleep furiously. ... 2

If we go back to primary school, where useful knowledge is sometimes imparted, we shall encounter the definition of a sentence as "the expression of a single idea". In these terms (1) (above) is a sentence and (2) is not, because (1) does impart an idea and (2) does not.

If at some time in the future, we find some way of attaching a meaning to (2), it will then become a sentence (in my sense). This will have to come about through some shift or extension in the meanings of its component words, not because of a shift in the grammar. Similarly, we can nowadays speak of splitting atoms. In Shakespeare's tine this was not possible: atoms were by definition indivisible, and so "atom" could not be the direct object of "split". We have a different conception of atoms, "atom" has changed its meaning, and we can now conjoin the previously incompatible.

From the above we see that the meaning of an utterance lies neither in its logical form, nor in the lambda calculus nor in set theory. Its meaning lies in the relation to the real world, or at least to our understanding of the real world. It follows that in analysing language we must take account of the world.

If we look about for a suitable basis for such an analysis then the fields of Artificial Intelligence (AI) and robotics are obvious candidates. Many rude things have been said about AI as it is currently practised, but the practitioners are at least trying to deal with out central problem. Robots, however, have one major advantage over us with our problem: they actually have a real world. A robot building a car can compare its instructions with the assembly line before it and decide whether those instructions are meaningful and, if so, what their meaning is. In translation we have only a text, a lexicon (however elaborate) and some rules of grammar. The robot can discard the operator's mutterings about his mother-in-law (the operator's - not the robot's: ambiguity is discussed later) as being irrelevant to the job in hand. We have no way of making that determination.

We are now in a position to say something about what goes into the lexicon and why. The lexicon, apart from purely morphological data, contains the relationship of words to each other. In this respect it mirrors the real-world relationships of the entities denoted by its entries. Where we have a lexical insertion rule, such as "eat" requires - or at least strongly prefers - an animate subject, this is a reflection of what "eat" means in the world. The very grammatical categories of words become reflections of their originals in the world; the words are no more than shadows and the rules of grammar are shadows of these shadows.

The problem of building a lexicon is the problem of building a model of the world, or of that part of the world in which we are interested. Many of the rules of subcategorization and lexical insertion are no longer grammatical but descriptions of external relationships reflected in the language being processed.

As the role of the lexicon changes from being primarily a word-list to being a world-model, so will the techniques that we use to implement it change. It now looks more like a data-base than a simple structured file, and the appropriate implementation is probably an entity-relationship or a network data-base. Strangely, a relational data-base would not be appropriate.

MEANING AND INTERPRETATION

In "Situation Semantics" Barwise and Perry discuss at some length the distinction between meaning and interpretation. They define meaning as being a property of a proposition ("sentence" for our purposes), and interpretation as being a property of an utterance, that is, a sentence used in a particular situation. The meaning of a sentence is what it can say about the world; its interpretation is what is does say. So the meaning of

The cat sat on the mat ... 3

has to do with every possible cat on every possible mat. Its interpretation, when I use (3), has to do with my cat sitting on some particular mat and cuddled up to my kitchen radiator.

For our purposes, meaning (of words) is what goes into the lexicon; interpretation is what is in the context that we are carrying around during processing. Alternatively, meaning is a permanent interpretation and interpretation is transient meaning. If we restrict ourselves to some particular area of discourse then we may be able to promote some

aspects of interpretation to the level of meaning, or to excise some parts of meaning.

A sophisticated translator would always remember all interpretations and, if the same ones cropped up often enough, would amend its lexicon appropriately. The amounts of both storage and processing require for such a procedure would be large, but with the current speeds of micro-processors and 300Mb discs available very cheaply, these are not insuperable problems.

It is important not to enquire too closely into just what the meaning of a term is. We can really only consider relationships between terms, and rely on their being based on real-world entities. Our theory may well lack a sound theoretical base, but so long as the universe continues to exist and the laws of physics do not change too drastically we are fairly safe. (Do not adjust your brain, reality is on the blink). If we try to define meanings too precisely we are liable to disappear in a mass of ontological gobbledygook and make very little progress.

[A semantic, or etymological curiosity: isn't it odd that we no longer expect teachers to be pedantic?].

HEADS AND MODIFIERS

> They've a temper, some of them - particularly verbs: they're the proudest - adjectives you can do anything with, but not verbs - however, I can manage the whole lot of them! Impenetrability! That's what I say.
>
> H. Dumpty

The notion of head and modifier can express a surprisingly large part of the grammar of a language. The head/modifier rule is

X' -> Y X ... 4a

X' -> X Y ... 4b

X' -> Y X Z ... 4c

where X, Y and Z are any grammatical categories, and X' is the generalization of X. Most languages tend to use just one of the variants of (4) over the others. Some languages use one variant exclusively. English uses (4a) usually and (4b) occasionally. Very few languages use (4c), the infix rule (I believe that Finnish does).

What (4a) says is that it is the last component of a word or phrase that determines its category. For example:

```
X is N, Y is A, X' is noun phrase
X is A, Y is Adv, X' is adjectival phrase
X is "ly", Y is A, X' is Adv
X is "ic", Y is N, X' is A
```

or:

```
N' -> A N          (tall stool)
A' -> Adv A          (very big)
Adv -> A "ly"        (quick ly)
A -> N "ic"            (bas ic)
```

This regards "ly" as carrying the 'adverb' category but no semantic weight, "ic" carries the 'adjective' category but no semantic weight.

A generalization of (4) is that nothing may come between a head and its modifier except another modifier of that same head. The choices of a variant of (4) is merely a choice of preferred ordering. Note that as a purely syntactic rule (4) is highly ambiguous for multiple modifiers. In

Pretty little girls' camp ... 5

we might mean

a camp for pretty little girls	(pretty is Adj)	...	5a
a camp for pretty-little girls	(pretty is Adv)	...	5b
a pretty camp for little girls		...	5c
a pretty little camp for girls	(pretty is Adj)	...	5d
a pretty-little camp for girls	(pretty is Adv)	...	5a

If the phrase is spoken rather than written then "camp" can be interpreted as a verb with "pretty little girls" as its subject-modifier, and with two interpretations of "pretty". In practice there is very little ambiguity in phrases of this kind. We already know from the context what is being spoken about, and the syntax rule (4) merely fills in some detail of what is being said about it. If there is any residual ambiguity it probably does not matter, or is deliberate (if you sent some toothpaste across London, would it have to go by tube?).

Probably most sentences that are ambiguous, even in context, involve adverbs. For example:

Smoking can seriously damage your health. ... 6

If "seriously" applies to "can" it means

Seriously, smoking can damage your health. ... 6a

If "seriously" applies to "damage" it means

Smoking can cause serious damage to your health. ... 6b

A few years ago

Hopefully the train will arrive on time ... 7

meant that the train would arrive on time and in a hopeful state of mind. Nowadays "hopefully" refers not to the train but to the speaker of the sentence.

So far I have considered only the analysis of a supplied utterance; I now come to the problem of generation. The head/modifier rule for English says that a modifier should precede its head and must be next to it. What happens if there are multiple modifiers; only one of them can occupy the favoured place? To get out of this hole we have rules of precedence, some of which are:

- Adjectives take precedence over articles and possessives:

  "the red book"

  only Shakespeare is allowed to write

  "good my lord".

- Strings of adjectives are dealt with according to a rather strange property. The distinction between nouns and adjectives is rather fuzzy; for example

  * The box is made of red 8
  The red box
  The box is made of plastic
  The plastic box
  ? The plastic red box
  The red plastic box

  Adjectives are ordered so that the less noun-like and more adjective-like come early, and they increase in noun-likeness up to the head of the noun phrase. The next word must be one that is not a noun, or is less noun-like, and so marks the end of the noun phrase.

- With nouns around a verb:

  - The subject takes precedence over everything (4a), except sometimes adverbs.
  - The direct object, which like the subject is not case-marked, goes behind the verb (4b).
  - All the case-marked indirect objects tag on behind (4b).

This is the rule for a VSO language like English. If we have a language which uses (4a) exclusively then we get an CSV syntax and the object precedes the subject. In this case either the imperative must differ inflectionally from the indicative or the object must be case-marked, otherwise we could not distinguish between an imperative with a direct object and an indicative without one. The Romance languages evade the problem by placing object pronouns before the finite verb but after the non-finite verb. There are not many nouns, as opposed to pronouns, that can be both subject of the indicative and object of the imperative verb, so that problem can be ignored with reasonable safety for the speakers of the language.
Languages that use (4b) exclusively, VSO, are known e.g. Welsh and Japanese.

Adverbs in English are most accommodating. They permit themselves to be push into any convenient slot, so long as they do not get too far away from their verb. Fowler under Position of Adverbs has a marvelous chamber of horrors, and some examples where he recommends the split infinitive.

> Such gentlemen are powerless to correctly ... 9
> analyse agricultural problems.
> A body of employers which still has power
> to greatly influence opinion.

In general adverbs go where they will cause the least disruption to the rest of the modifiers. With an intransitive verb they follow and with a transitive verb they precede the verb. For an infinitive with a direct object the adverb either splits the infinitive or follows the object. About the only time that an infinitive cannot be split is when it is itself a modifier of a preceding participle; in this case some other convenient slot has to be found for the adverb.

In all choices of word-ordering, long modifier phrases tend to get pushed outwards so as to keep the nucleus of the sentence as compact as possible.

If we return to the boldly split infinitive of (1) then we could, but need not, rephrase it as

> To split infinitives boldly where no man ... 10 has
> boldly split before.

Note that the second "boldly" ought not to be moved. The preferred place for an adverb is after the verb, "has split boldly before", but with another adverb there the adverbs can both be next to the verb by moving one back to between "has" and "split". If we go back to the original of this sentence

To boldly go where no man has gone before ... 11

then, in the absence of a direct object, there is no reason to split the infinitive. The adverb should have remained in its normal place after the verb, writing

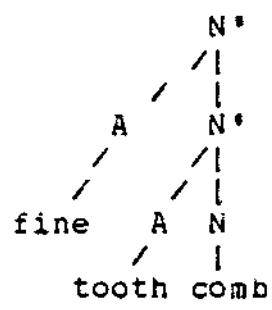
To go boldly where no man has gone before ... lla

so that our teeth are no longer set on edge.

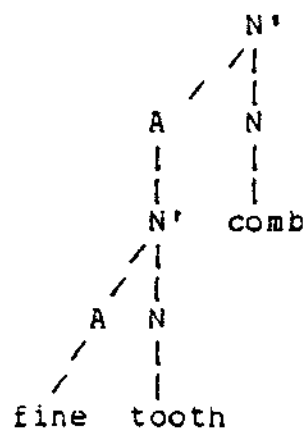
COMPOUNDS

> Well "outgribing" is something between bellowing and whistling, with a kind of sneeze in the middle.

A feature of English is that almost any sequence of words, however unrelated, that occurs sufficiently often can become a unitary compound. An example is the recent vogue phrase "fine-tooth comb" that has become "fine tooth-comb" - whatever a "tooth-comb" might be! Effectively, people analyze the phrase as two adjectives plus a noun:



rather than as adjective+noun acting as adjective, and noun:



using the usual vagueness between nouns and adjectives.

I believe that "how to" is now a compound adverb, consequently the title of this paper does not contain a split infinitive, whatever it may have done 20 years ago.

CONCLUSIONS

> In order to more precisely define the relationships derived by the procedure and to motivate the method used, ...
>
> A technical report

Language is simple; it is the world that is complicated and it is this complication in the world that is reflected in our highly efficient language structures. The problem of achieving high quality machine translation is the problem of modelling the world and our changing view of it. What I am advocating is that we simplify our linguistics and cheat wildly on the modelling.

REFERENCES

> The original question "Can machines think?" I believe to be too meaningless to deserve discussion. Nevertheless, I believe that as the end of the century the use of words and general educated opinion will have altered so that one will be able to speak of machines thinking without expecting to be contradicted.
>
> Alan Turing
> "Computing Machinery and Intelligence"
> **Mind,** Vol LIX. No 236 (1950)

HUDSON, R. A. (1974): **Word Grammar**, Blackwell

CHURCH, K. and PATIL, R. (1982): "Coping with Syntactic Ambiguities", MIT

BACH and HARMS: **Universals in Linguistic Theory**

FRAZIER and FODOR (1978): "The Sausage Machine", **Cognition** Vol.6

HOELESTRA et al: **Lexical Grammar**

JACKENDORFF, Ray: "How to Keep Ninety from Rising", **Linguistic Inquiry**, Vol.10.1 pp 172-176