[*Proceedings of a Workshop on Machine Translation*, July 1990, UMIST]

Large-scale Discourse Structures

and

Machine Translation

Karen SPARCK JONES
Computer Laboratory
University of Cambridge
New Museums Site
Pembroke Street
Cambridge CB2 3QG, UK

May, 1990

Large-scale discourse structure and MT

Karen Sparck Jones, May 1990

There appears to be large-scale structure in discourse (dialogue or single-source). This structure, which may be called argument or rhetorical structure, seems to be as much semantic as syntactic. Its specific character is not at all well understood; there may be several distinct categories of structure. But assuming that, as exemplified by such examples as 'claim, counterclaim, modified claim, ...' or 'definition, elaboration, illustration ...', it exists, is it necessary to take it into account for MT?

The low level reason - call it the accuracy reason - for looking at structure outside the sentence is to deal with problems like pronoun resolution when gender choices have to be made, or reference determination when articles have to be supplied. But it is not clear, setting aside the extent to which full language understanding would be needed to deal with these problems, whether it is necessary to know, say, that S2 stands in elaboration relation to S1, to control pronoun resolution: it could be sufficient to use local syntactic structure and cohesion mechanisms.

The high-level reason - call it the fidelity reason - is that we need to recognise this structure to ensure that it is preserved during translation. There could be two different reasons for this. One is that it might be needed to disambiguate functional connectives like "Thus", or to preserve their effect in local translation where there are no simple equivalents. The other, more interesting one is in relation to interlingual approaches to translation where local expressive fidelity is not sought but where source content is to be preserved. Argument structure may be an important component of discourse content (eg in a scientific paper or abstract), and this could be completely and damagingly lost in approaches focusing on eg domain frames absorbing the content of many individual source sentences, but treating this sentential content in a straightforward way.

large scale discourse structure

needed for MT?

local structure needed

anaphors, cohesion

large scale structure needed for

summarising

eg intro, problem, method ...

first day, second day ....

proposal, pro, contra...

large scale structure :

nonlinguistic / linguistic

bottom up / top down

nonlinguistic

bottomup

eg generalisation hierarchy

top down

eg script

linguistic

bottom up

eg focus space

top down

intermediate

rhetorical relations
rhetorical schemata
argument grammar

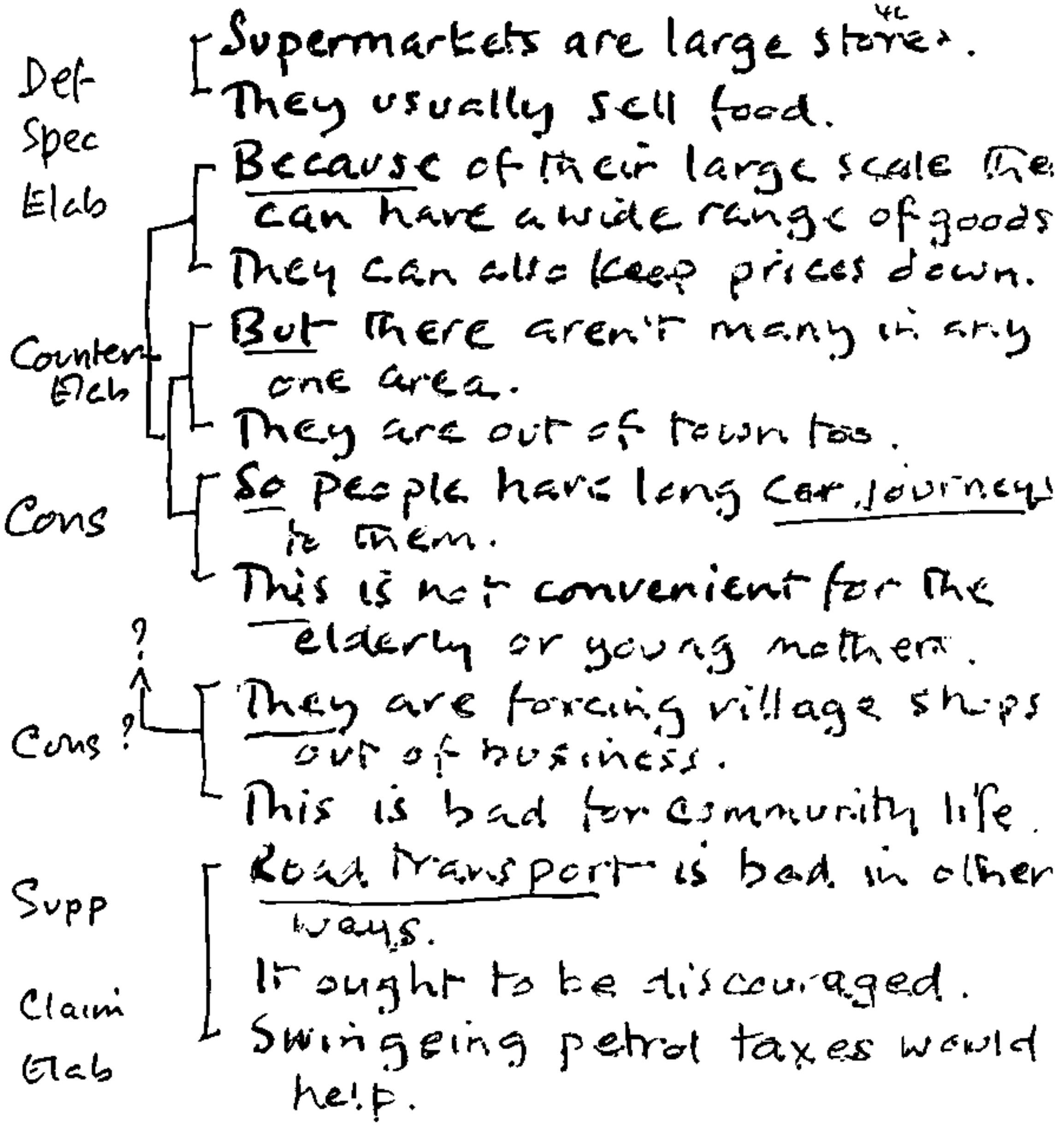
form eg analogy, comparison
content eg cause, constituency
function eg counterargument,
evidence

utilise in MT for

interpretation

Consistency ?

EG



Def — Supermarkets are large stores.
Spec — They usually sell food.
Elab — Because of their large scale they can have a wide range of goods
They can also keep prices down.
Counter Elab — But there aren't many in any one area.
They are out of town too.
Cons — So people have long car journeys to them.
This is not convenient for the elderly or young mothers.
?
Cons ? — They are forcing village shops out of business.
This is bad for community life.
Supp — Road transport is bad in other ways.
Claim — It ought to be discouraged.
Elab — Swingeing petrol taxes would help.

The labels and brackets both apply directly to the source text

How far beyond sentence to go
for accuracy
fidelity in MT?

Can it be done with
linguistic apparatus ?
eg text grammar