Anticipating the Reader's Problems and the Automatic Generation of Paraphrases

Nils Lenke

Gerhard-Mercator-Universitaet-GH Duisburg FB3 - Computerlinguistik
Lotharstr. 65, D-47048 Duisburg
voice: +49 (0)203-379-2007; e-mail: he233le@unidui.uni-duisburg.de

0. ABSTRACT

The notion of paraphrase is discussed and compared with the similar notion of periphrase. The role of paraphrases in oral communication is described, and the results of a study on the role of paraphrases in texts are given. Finally, a system which models the use of paraphrases in texts is described.

1. PARAPHRASES IN DIALOGUES

If you look at ordinary dialogues you will find that communication failures - i.e. different types of misunderstandings - happen frequently, cf [Ringle & Bruce 1982]. One important technique for the participants of the communication to solve these problems is *paraphrasing*, that is, saying it again in other words. Paraphrases can be offered by the hearer ("Is it this what you want to say: ...") or requested from the speaker by the hearer ("Huh? I don't understand."). These kind of paraphrases may be called *communicative* or *pragmatic* paraphrases.

2. OTHER NOTIONS OF "PARAPHRASE"

Notions of "paraphrase" exist which differ from the one presented above. In linguistics, especially Transformational Grammar, cf. e.g. (Smaby 1971), (Nolan 1970), the paraphrase relation is induced by the rules of the language system. Two formulations count as paraphrases of each other if they can be derived from a common deep structure, e.g. the active and the passive version of a sentence. So, the paraphrase relation is completely independent of the situation and communication participants. This view has been heavily criticised, cf. (Ungeheuer 1969).

In CL, the generation of a surface form from a meaning representation is sometimes called paraphrase generation, especially if different surface forms can be generated for the same meaning representation. An ambiguity exists here, because the paraphrase relation can be meant to hold (a) between the meaning representation and the NL text derived from it or (b) between two alternative formulations which could both be derived from the meaning representation.

I will simply call case a) "generation" because that is what it means: deriving a text from an underlying meaning representation. Case b), exemplified by (Goldman 1975) and most work in the area of Meaning-Text Models, cf. e.g. (Iordanskaja, Kittredge & Polguère 1991), (Mel'cuk 1981), stresses the possibility of an alternative formulation which could be uttered *instead* of another formulation, whereas in section 1 we talked of paraphrasing as uttering a formulation *in addition* to another formulation. To differentiate between these cases I will not call case b) *paraphrase* but - in accordance with classical rhetoric - *periphrase*.

3. RELATED WORK IN CL

Quite a lot of work exists on the use of paraphrases in connection with NL-database frontends, cf. e.g. (McKeown 1979), (Meteer & Shaked 1988). The formal representation gained from the user's query is translated back to NL again and the user is requested to indicate if the system understood him correctly. This fits nicely into the framework from section 1.

As indicated above, much of the work presented under the title "paraphrase generation" should better be called "periphrase generation". Reiter's (1990) system FN generates - depending on the user model entry for the problematic word "shark"- one of the following alternative formulations:

1a) There is a shark in the water
1b) There is a dangerous fish in the water

Similarly, the system WISBER (Horacek 1990) generates one of the following formulations, where the problematic word is "Notgroschen" (rainy day fund):

2a) Haben Sie einen Notgroschen? [Do you have a rainy day fund?]
2b) Haben Sie ein Sparbuch mit zwei Nettomonatseinkommen? [Do you have a savings account with two month's net income?]

In the terminology advocated here, the b)-cases are *periphrases* of the a)-cases. Real formulations with paraphrases would look something like this:

1c) There is a shark,that is, a dangerous fish, in the water
2c) Haben Sie einen Notgroschen, d.h. ein Sparbuch mit zwei Nettomonatseinkommen?
[Do you have a rainy day fund, that is, a a savings account with two month's net income?

It will be discussed below under which circumstances such utterances could be superior to the a)- or b)-cases.

4. ANTICIPATION OF MISUNDERSTANDINGS AND THEIR AVOIDANCE

Turning now to the generation of written texts it seems to be a bit paradox to do this in connection with paraphrases, since in section 1 we showed them to be a phenomenon of *dialogue*, i.e. oral communication. But paraphrases do play a role in texts as well, especially when *anticipation* is considered. This can already be noted in the case of spoken language. A well known model of the production of spoken language is the one of Levelt (1989). One of its main aspects is the existence of control and revision loops which can be used to monitor the planned or realized utterance and detect errors in it. So, part of the errors can

already be anticipated in advance by the speaker before the hearer even gets to hear the problematic utterance.

When we now turn to written language again, we also find the concept of problem anticipation and revision loops. These are of even greater importance here because the reader normally has no chance of signalling his problems with a text to the author. So, the author has to take the role of the reader and anticipate problems he might have with the text. Most models of the writing process thus include a revision loop, cf. the well-known model of Hayes and Flower (1980). In CL, this mechanism is known under the name *anticipation-feedback loop*, cf. (Jameson & Wahlster 1982), and in the form of revision-based generation systems, cf. (Gabriel 1988), (Vaughan & McDonald 1986).

What are the options for an author if he detects trouble sources in his planned text? He may choose to

a) add a meta-comment; the addition of meta-comments (Sigurd 87) like "loosely speaking", "to say it frankly", "a kind of", etc. is often used to indicate to the reader how to interpret a problematic utterance.
b) add a further, alternative formulation (a paraphrase) or
c) replan the text (formulate a periphrase).

The rest of the paper will solely deal with b) and c). What was said so far leads to the following hypothesis:

Writers of texts anticipate reader problems, and, in some cases, include paraphrases to avoid these troubles.

5. A STUDY ON PARAPHRASES IN TEXTS

A study, cf. (Lenke, in preparation) for details, was conducted in order to find occurrences of paraphrases in texts and analyse them with the aim of checking the hypothesis mentioned at the end of section 4.

First, a small corpus of German texts was scanned manually for paraphrases; the major results were:

- Paraphrases of the kind described above can indeed be found. Typical examples of such paraphrases are[1]:

(3) "... introduces the notion of **multiple inheritance** - that is, the ability of a class to have more than one direct base class - and presents ..." [p. 182]

(4) "A language is said to support a style of programming if it provides facilities that make it **convenient** (reasonably easy, safe, and efficient) to use that style." [p. 14]

- only part (roughly 50%) of the paraphrases are announced by indicators like "that is", "in other words", parentheses or hyphenation. The other paraphrases are simply added as an apposition to the paraphrased term.

- the total number of paraphrases differs vastly between text types: in narrative texts few and mostly unannounced paraphrases occur; in more technical texts, especially manuals and introductory texts, many paraphrases.

In the second phase of the study, the LIMAS corpus of German (1 million running words from 500 texts of different types) was then scanned automatically for the most common German paraphrase indicators (a.o. "d.h.", "das heißt", in anderen Worten", "also") Well above 1000 occurrences of paraphrases were found and analysed. The results of the first phase could be confirmed. Other results were:

- the syntactic form of the paraphrases is in most cases either a complete sentence (in which another complete sentence is paraphrased) or an apposition, which belongs to the same syntactic category as the word/phrase it belongs to.

- Paraphrases are directed to quite different problem sources which were anticipated by the author. Among the different types found were the following:

1. problematic lexical items
a) unknown words (cf. examples 3 above)
b) ambiguous words;
c) words of abstract nature which obtain their concrete meaning through the context in which they occur. The paraphrases indicate the direction in which this concrete meaning should be searched. Cf. example 4 above.

2. reference problems
a) ambiguous anaphoric references, e.g. pronouns;
b) anaphoric expressions where the referent is very distant (causing memory problems)
c) missing knowledge to understand referring expression.

3. problems induced by rhetoric figures (metaphors, metonymy).

4. inference problems
a) problems of aspectualization. (only some aspects of the meaning of a word are relevant in a certain context).
b) problems of logical inferences. (Obvious and relevant inferences from an utterance might be too difficult to draw by the intended reader).

Thus, one can conclude that paraphrases are indeed used by authors to avoid anticipated reader problems. These problems can be of all those types that have since long been noticed in the area of NL understanding.

6. IMPLEMENTED MODEL

The next step in the project was to design and implement a model which describes this use of paraphrases in texts. It should answer the following questions:

- How can problems of the reader be anticipated?
- Under which circumstances are paraphrases the adequate answer to this problems (and not, say, periphrases or meta-comments)?
- How can paraphrases be generated?

Three well known approaches to NL generation are combined in the model : user modelling, anticipation-feedback

[1] the following English examples all stem from [Stroustrup 1991] and were collected just to be English examples suitable for the presentation in this paper.

loops and revision-based text generation. Its architecture is shown in Fig 1:

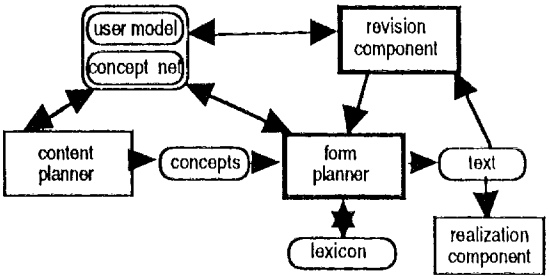

Fig 1: the system's architecture

The main feature is the revision/ anticipation-feedback loop, which is highlighted in the figure.

The types of problems for which paraphrases can be generated by the system are restricted to problems which occur during lexicalization and involve only conceptual knowledge (no assertional knowledge) in order to restrict complexity. These are (in terms of section 5) the types 1a, 1b, 3a, and - with restrictions - 4a, which are (together with type 1c) by far the most frequent types occurring in natural texts. The other types could principally be dealt with in a similar fashion. A corpus of about 25 examples, all collected from the same source, the manual for the Apple Macintosh operating system 7, were used as a basis. The advantages of this approach is that all examples are based on a common domain (knowledge about Macintosh computers), so that a common lexicon and a common knowledge base can be used for all of them. Of course, the techniques and principles used are not restricted to this set of examples and could be transferred to other domains.

6.1 An example

To demonstrate how the components of the systems work together consider example 7, from the corpus on which the system is based:

(5) Alle Macintosh Modelle sind mit einem **Steckplatz** oder *Anschluß* für Geräte ausgestattet, der die *SCSI-Schnittstelle* (Small Computer System Interface) unterstützt. [all Macintosh models are equipped with a **slot** or *interface* for SCSI-devices]

The content planner of the system is only implemented as an *oracle*, that is, it is preset to produce the concepts to be formulated and to answer certain questions by the form planner as if it were a full-fledged content planner in a complete NL system. In the concrete example, it would first inform the other components that the linguistic context of the target item consists of the concept *Macintosh* (the only concept that precedes *slot* in the planned sentence) and would then request the form planner to verbalize the concept *slot*.

The form planner would then look up the first possible linguistic items for the concept *slot* in the lexicon. The lexicon not only incorporates information about the linguistic items but also about their connections to items of the concept-base. These connections take the form of ZOOM-schemata, as known from the WISBER system, cf. (Horacek 1990). Briefly, ZOOMs are links between concepts or small sub-structures of the concept-network on the one hand and linguistic items (words) on the other hand.
In our example, the first choice to verbalize *slot* would be 'Anschluß'. This proposal is then put forward to the revision component which tries to anticipate reader trouble. To do this, it uses a simple user model, which employs the well known stereotype approach (Wahlster & Kobsa 1989). All concepts, lexical entries and ZOOMs belong to one of the three categories *common vocabulary*, *computer jargon* and *Macintosh specific jargon*. The static part of the user model then simply consists of three variables which indicate if the intended reader is expected to be familiar with the respective jargon.

This user models differs from other approaches because it allows the special value "?" which indicates incomplete (you never know all about the readers) or inconsistent (a text can be meant simultaneously for novices and experts) knowledge. From this static part of the user model a default value can be calculated which can be overridden through learning (see below). To be a bit more exact, *two* values are calculated in a kind of "worst-case-analysis" due to the "?" values in the user model.
In our example two ZOOM-schemata exist for slot:

slot <-> 'Anschluß'
slot <-> 'Steckplatz'

'Anschluß' (and the ZOOM connecting it with slot) is marked *Macintosh*, the alternative lexical entry 'Steckplatz' is marked *common*. So, if the user model indicated that Macintosh vocabulary was *yes*, the revision component would judge the wording 'Anschluß' ok and the realization component would output

```
"Alle Macintosh Modelle sind mit EINEM
ANSCHLUSS für Geräte ausgestattet, der
die SCSI-Schnittstelle unterstützt."
```

But now consider a user model which indicates that the knowledge of computer and Macintosh jargon is known to be *no*. Of course, the revision component would indicate that the term 'Anschluß' cannot be used. A possible solution would be to generate a periphrase, i.e. replacing 'Anschluss' by 'Steckplatz' which would be the next choice of the form planner. This would then be accepted by the revision component. In some cases, however, this would be less than perfect: (a) if the concept has repeatedly to be verbalized in the course of the text, (b) if there are stylistic reasons to use the first choice term (here: 'Anschluss'), (c) if there are pedagogical reasons to use the first choice.

(a) consider a case in which the periphrase is a longish definition. It would be a bore to replace a short term by this definition 15 times around the text. So you do it once and simply use the now learned term in the rest of the text.
(b) Certain texts can loose their "feel" if stripped of e.g. the expert vocabulary of a certain area.
(c) Manuals and introductory texts are often meant to teach the vocabulary in addition to the concepts. In this case it would be nonsense to replace the to-be-taught vocabulary by easier "terms".

All these conditions can only be determined by the content planner (demonstrating the need for an interaction between form planner and content planner); in the system, the form planner asks the content planner, which works as an oracle, i.e. gives the correct answers (by forwarding the questions to the human operator). If one of the conditions holds, it would be unwise to formulate a periphrase. The next choice of the form planner would then be to ask the content planner to completely replan this part of the text, namely to include a new sentence defining the problematic term. The system output looks like this:

```
ANSCHLUSS BEDEUTET STECKPLATZ. Alle
Macintosh Modelle sind mit EINEM AN-
SCHLUSS für Geräte ausgestattet, der
die SCSI-Schnittstelle unterstützt.
```

Even this solution doesn't work in some cases and that is where paraphrases come into play. If stylistic variation is necessary or if the problematic term is embedded in the definition of still another term it is the right place to use a paraphrase:

```
Alle Macintosh Modelle sind mit EINEM
ANSCHLUSS D.H. EINEM STECKPLATZ für
Geräte ausgestattet, der die SCSI-
Schnittstelle unterstützt.
```

6.2 A second example

Just another path may lead to the generation of paraphrases for an unknown term, as the next example will show:

(6) "Mit der *Maus* - dem Gerät zum Zeigen und Klicken - werden die meisten Macintosh Funktionen aktiviert."
[With the mouse - the device for pointing and clicking - most Macintosh functions are activated]

Fig. 2 shows the part of the conceptual network underlying this example:

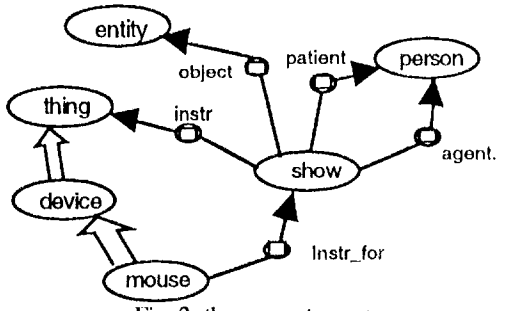

Fig. 2: the concept *mouse*

The term 'Maus' is classified *computer jargon* and may not be known to the user. The replacement of the term by a definition (no synonym is available) yields the danger of encouraging false conversational implicatures by the reader, cf. (Reiter 90). Consider a user model where *computer* and *Mac jargon* are indicated as "?". A worst case analysis by the revision component would show that the use of 'Maus' is inappropriate because some novices wouldn't know the term, but that the periphrase 'Gerät zum Zeigen (und Klicken)' is inappropriate either, because some experts will know the term 'Maus' and conclude from its absence that some other pointing device, but not the mouse, was meant.[2] So, a paraphrase would again be the best solution. The system thus generates:

```
Mit DER MAUS D.H. DEM GERAET DES
COMPUTERSYSTEMS ZUM ZEIGEN werden die
meisten Macintosh Funktionen akti-
viert.
```

Here the paraphrase is a definition of the form *per genus proximum et differentia specifica* which results from part of the systems' concept net shown in figure 2. The system is capable of generating two other forms of definitions (paraphrases), definition by antonymy and by enumeration.

6.3 Detection and resolution of ambiguity

Up to now, only the problem type of unknown words has been discussed. Due to lack of space only one more problem type which leads to the generation of paraphrases can be discussed, namely the problem of *ambiguous words*. This problem type has since long been discussed in the area of NL understanding. Techniques for its solution include the use of spreading-activation mechanisms working on conceptual networks, cf. (Hirst 1987). This can now be used for the purpose of problem anticipation. We just try to disambiguate terms and interest ourself in the cases in which it fails: these are candidates for paraphrase generation. Cf. the following example from the corpus:

(7) "Das aktive Fenster steht **im Vordergrund**[,] also vor allen anderen geöffneten Fenstern."
[The active window stands in the foreground, that is, in front of all other open windows]

Here, for beginners two readings of 'im Vordergrund' are possible: a literal (this is the correct reading) and a metaphorical (in the sense of "important, to be regarded") which are equally propable. The revision component comes to this conclusion by conducting a worst case analysis using the concept net, an activation-spreading algorithm and the user model. Only those concepts and links that are known to a reader may forward energy, so in the case of "?" values in the user model, both alternatives have to be tested (hence the term "worst case anaysis"). If comparable quantities of the activation energy induced into the net by the linguistic context find their way to both (or more) readings (concepts) of the ambiguous terms it is concluded (and then indicated to the form planner by the revision component) that the ambiguity might not be resolved by the reader. Then, a paraphrase could eventually (in a process similar to that described above) be generated, defining the correct reading. See (Lenke, in preparation) for details of the spreading-activation mechanism used.

---

[2] cf. Reiter's (1990) "dangerous fish" vs. "shark" example.

6.4 Two more features of the system

These can only be discussed briefly. See (Lenke, in preparation) for details.

- Paraphrases of the aspectualization type (see above, section 5) can also be generated. Here, only one of the defining elements of a concept, either the superclass (genus proximum) or one of the roles (differentiae) is verbalized. At the moment, this kind of paraphrase is only generated when requested by the content planner; in the future, it will be necessary to model the anticipation of inference processes based on relevance by the reader to correctly predict the need for such paraphrases. An example from the corpus, the underlying concept net and the equivalent produced by the system are shown below.

(8) Durch das Klicken **werden die Objekte *aktiviert***, d.h., sie werden nun schwarz (oder in einer anderen Farbe) dargestellt und somit hervorgehoben. [Caused by the clicking the objects are activated, that is, printed in black (or another colour) and so highlighted]

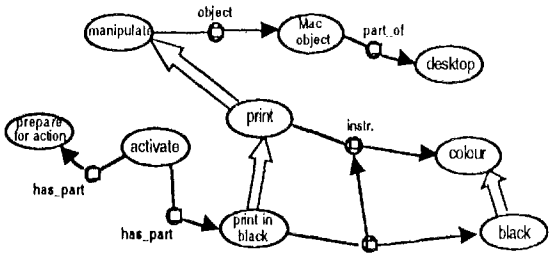

Fig. 3: the concept *activate*

```
Durch das Klicken werden die Objekte
AKTIVIERT D.H. SCHWARZ DARGESTELLT .
```

- if a paraphrase for an unknown term has been generated, it can be concluded that the reader now knows this term. This is modelled by an active component of the user model which overrides the default values computed by the static component decribed above. So, only for the first (or first and second) appearence of a term a paraphrase is generated. Thereafter the term is simply used. This nicely mimics the observations made in naturally occurring texts.

7. IMPLEMENTATION DETAILS

The system is implemented in an object-oriented programming language and runs on Macintosh computers. It contains a conceptual network similar to KL-ONE, consisting of approx. 130 concepts and 65 roles. Its lexicon consists of 70 ZOOM schemata and 50 lexical entries.

8. FUTURE WORK

Some possibilites for future work have already been indicated in the text, most notably the embedding of the procedures described into a full-fledged NL-system. The approach described could also be transferred to other kinds of possible reader problems as enumerated in section 5. Since these are the problem areas of NL-understanding, algorithms exist which try to solve the understanding problems posed by these language features. These could be used to predict failure (as was demonstrated above for the activation-spreading mechanisms).

9. REFERENCES

Gabriel, R.P. (1988). Deliberate Writing. In: McDonald & Bole Eds. *Natural Language Generation Systems*. pp.1-46.

Goldman, N.M. (1975) Conceptual Generation. In: R.C. Schank Ed. *Conceptual Information Processing*.

Hayes, J. R. & L. S. Flower (1980). Identifying the Organization of Writing Processes. In L.W. Gregg & E.R. Steinberg Eds. *Cognitive Processes in Writing*. pp. 3 - 30.

Hirst, G. (1987) Semantic Interpretation and the Resolution of Ambiguity.

Horacek, H. (1990). The Architecture of a Generation Component in a Complete NL Dialogue System. In Dale, Mellish, Zock Eds. *Current Research in Natural Language Generation*. pp. 193-227.

Iordanskaja, L., R. Kittredge & Al. Polguère (1991). Lexical Selection and Paraphrase in a Meaning-Text Generation Model. In Paris, Swartout, Mann Eds. *Natural Language Generation in AI and CL*.

Jameson, A.& W.Wahlster (1982). User Modeling in Anaphora Generation: Ellipsis and Definite Description In: *Proc. of ECAI 82*. pp. 222-227.

McKeown, K. R. (1979) "Paraphrasing Using Given and New Information in a Question-Answer System", in Proc. of the 7th Conference of the ACL, La Jolla, 1979, pp. 67 - 72.

Lenke, N. (in prep.). Paraphrasen - Lösungen für antizipierte Leserprobleme bei der automatischen Textgenerierung. Dissertation, Univ. of Duisburg.

Levelt, W. (1989). Speaking. From Intention to Articulation.

Mel'cuk, I. A.. Meaning-Text Models (1982): A Recent Trend in Soviet Linguistics. *Ann. Rev. Anthropology* 10:27 -62.

Meteer, M. & V. Shaked (1988). Strategies for Effective Paraphrasing. *Proc. of Coling '88*. pp. 431 - 436.

Nolan, R. (1970). Foundations for an Adequate Criterion of Paraphrase.

Reiter, E. (1990). Generating Descriptions that Exploit a User's Domain Knowledge. In Dale, Mellish, Zock Eds. *Current Research in Natural Language Generation*. pp. 257 - 285

Ringle, M. H. & B. C. Bruce (1982). Conversation Failure In Lehnert; Ringle Eds. *Strategies for Natural Language Processing*. pp. 203- 221.

Sigurd, B. (1987). Metacomments in Text Generation. In G. Kempen Ed. *Natural Language Generation*. pp. 453-461.

Smaby, R. M. (1971). Paraphrase Grammars.

Stroustrup, B. (1991) The C++ Programming Language. 2d. Ed.

Ungeheuer, G. (1969) Paraphrase und syntaktische Tiefenstruktur. *Folia Ling*. 3. pp. 178-227.

Vaughan, M. & D. McDonald (1986). A Model of Revision in Natural Language Generation. *Proc. of the 24th Annual Meeting of the ACL*. pp. 90-96.

Wahlster, W. & A. Kobsa (1989) User Models in Dialog Systems. In A. Kobsa & W. WahlsterEds. *User Models in Dialog Systems*. pp. 4 - 34.