AMBIGUITY AND AUTOMATED CONTENT ANALYSIS.

Drs. M. Boot.

1. Introduction.

"Content analysis" is a form of language analysis and "language analysis" usually is a form of linguistics. The word "automated" refers to the computer as a tool in any kind of human activity. Nevertheless, the word "automated content analysis" however does not refer to computational linguistics.
When studying the classical work on automated content analysis (Stone, 1) one will not read about the topics discussed for example in the classical work on computational linguistics (Hays, 2). The same holds for a comparable area of scientific investigation concerning language, that part of "artificial intelligence" dealing with "understanding natural language" (Winograd, 3) called natural language dialog simulation. On the contrary, "automated content analysis" seems to be a topic in social sciences and simulation of communicating in dialogues seems to be a topic in psychology, as can be concluded from the furious controversy between the artificial intelligence groups and the professional linguists (Schank e.a., 4). This paper will try to bridge the gap between computational linguistics and automated content analysis. Concerning the phenomenon of ambiguity, the paper will provide new proposals which are based on very old and recently very much neglected insights of linguistics. These proposals have lead to the construction of a model for automated disambiguation. The first results will be shown of a computer program which is based on that model.

2. What kind of language analysis is content analysis?

2.1. Content Analysis as Textanalysis.

Already from the pages overviewing the contents of the book on the general inquirer (Stone, 1) it is clear that content analysis means text analysis. It is important to state this fact, because linguistics is primarily concerned with sentence analysis. All problems concerning the questions about pieces of language bigger than a single sentence are indeed neglected by pure linguistics. Those questions are tackled in

such different scientific endeavours as stilistics, philosophy and rhetoric. In German recently the first steps are set to define a new branch of linguistics called "Textlinguistik" (Kalmeyer, 5). Textlinguistik, however, is more concerned with theoretical discussions than with the problems of concrete text analysis.
Hence, it appears that concrete text analysis should be based on models derived from diverse and widely scattered human knowledge about texts. Consequently, the reader will not be surprised to discover that text analysis is as yet not a very well developed field of scientific inquiry. Of course this results in a great challenge to workers in the field.

2.1.1. Texts: Non-fiction.

Of course a definition like "text analysis" is to wide: it is evident that it may not be very prudent to handle together such diverse texts as a funeral sermon, a love letter, a Petrarca sonnet, or some coverage of a sport event. Therefore the field of analysis should be narrowed down.
Indeed practice narrowed down the broad field of content analysis to a narrower field that could be indicated more or less by the way of looking at texts by the text analyser. Content analysis is concerned with a specific function of texts: the social function of communicating thoughts, emotions or so called objective information.

This indication of the field makes disappear all those kinds of texts studied by the science of literature, because of the fact that this science is primarily concerned with a "deeper" level of communication than the type of information included in normal, the so called non-fiction texts.
By now we narrowed down the field of investigation to the type of text that could be indicated as "informational" text. For this paper however this typology is not precise enough. To demonstrate that we must first focus on the different aspects of the contents that can be communicated by texts of the type and in the function we want to study, let us consider the following text:
"Mr Chief Justice, President Eisenhower, Vice President Nixon, President Truman, reverend clergy, fellow citizens, we observe today not a

victory of party, but a celebration of freedom - symbolizing an end, as well as beginning - signifying renewal, as well as change. For I have sworn before you and Almighty God the same solemn oath our forebears prescribed nearly a century and three quarters ago.
The world is very different now. For man holds in his mortal hands the power to abolish all forms of human poverty and all forms of human life. And yet the same revolutionary beliefs for which our forebears fought are still at issue around the globe - the belief that the rights of man come not from the generosity of the state, but from the hand of God.
We dare not forget today that we are the heirs of that first revolution. Let the word go forth from this time and place, to friend and foe alike, that the torch has been passed to a new generation of Americans - born in this century, tempered by war, disciplined by a hard and bitter peace, proud of our ancient heritage - and unwilling to witness or permit the slow undoing of those human rights to which this Nation has always been committed, and to which we are committed today at home and around the world." (Leech: 6; 65).

Confronted with the question, what concrete information is communicated by this text, as by every text whatsoever, one has to be aware of the situation ("context") in which this text was produced. If the reader supposes to read a chapter from a scientific, say historical, book, he would wonder, how one should be able to "observe ... today ... a celebration of freedom". How to find the criterion to verify that "the world is very different now" and so on.
It might be evident that those kind of questions looking for verification procedures of the information communicated are out of place. Nobody, however, would read this text with this kind of criteria or if a person were reading that way one would think that he was misinterpreting the text.
This example demonstrates that it is indeed true that the situation must be known for which the text was created before content analysis will be feasible. With respect to the Kennedy speech Geoffrey Leech characterizes the situation as follows:

> "If we regard the main audience of the speech as that majority of "average Americans" who are emotionally committed to the institutions of their country, there is scarcely anything that can be

disagreed within the speech." (Leech: 6, 65)

The text is like a remark about the weather.

"The significant similarity between President Kennedy's address and a remark about the weather should not, of course, blind us to the emotive power of the speech, and to the use of political affective words (rights of man, human rights) which shows its affinity with political propaganda. But the function of the speech is not so much to change attitudes, as to reinforce or intensify them."(Leech: 6, 65-66).

The function of this text is the "function of maintaining cohesion within social groups" (Leech: 6, 62), it is known as the pathic function of language. In this case language does not use the informational function of communication in the real sense of that word.

On the basis of the foregoing we can define the text we want to study as the "informational text in that sense of the word". We wish to exclude the pathic function of communication, as we excluded the art function. This exclusion will enable us to define one of the most fundamental problems of ambiguity as far as computational linguistics is concerned.

From now on such very important functions of ambiguity as irony and methaphorical use of words are excluded from this study, because they are creative expressions of language functions being not "informative" in the restrictive way defined in this chapter. By doing so we arrive at a definition of the word text.

2.1.2. Conclusion: Definition of text.

A text is a coherent integration of informative statements encoded in written form.

3. Ambiguity, homonomy and homography.

As the basic unity of a text we defined the statement. A statement is a concept foreign to linguistic analysis. The basic unit of linguistic analysis which is comparable to a "statement" is the "sentence". It will be shown that these 2 concepts are not equivalent. These two units of analysis, however, have one characteristic in common: they are constructed by words. We shall describe a model to disambiguate words in texts and we will demonstrate how this model leads to the determination

of statements included in the text. Before we can come to a description of that model, however, we have to follow the previous procedure to define the word "text": we shall first exclude a special type of ambiguity on the word level. Let us consider the following text:

"A distinguished Negro sociologist tells of an incident in his adolescence when he was hitchhiking far from home in regions where negroes are hardly ever seen. He was befriended by an extremely kindly white couple who fed him and gave him a place to sleep in their home. However, they kept calling him "little nigger" - a fact which upset him profoundly even while he was grateful for their kindness. He finally got up courage to ask the man not to call him by that "insulting term".

'Who's insultin' you, son?' said the man.

'You are, sir - that name you're always calling me.'

'What name?'

'Uh ... you know.'

'I ain't callin' you no names, son.'

'I mean your calling me "nigger".'

'Well, what's insultin' about that? You are a nigger, ain't you?'

(Hayakawa: Language in Thought and Action, pp. 90-91)

This example shows that one and the same word can have more than one meaning. In this case those differences in meaning were affective. Of course, this kind of difference in the meaning of one single word is extremely important to communication and content analysis. Working with computers, however, this type of ambiguity can not yet be solved. Before resolving this kind of ambiguity, a more "basic" one has to be removed from the text. For this reason this study will exclude the kind of ambiguity, which is known as "homonymy" (Boot, 7).

The more basic type of ambiguity is illustrated by the following examples:

1. The professor is a mean old man.
2. I mean to go downtown tomorrow.

This example demonstrated how one and the same wordform ("mean") can constitute different words. In the first sentence "mean" has a meaning of "unkind", "vicious", "vindictive" or something like that. In the second sentence "mean" is a colloquial substitute for "intend". In a linguistic description those words are assigned to different "word classes": the word form "mean" in the first sentence is assigned to the class

"adjective". The word form in the second sentence is assigned to the word class "verb". In other words, linguistics provides for a well developed system to disambiguate this kind of difference in word meanings. The computer (not being a human being or a linguist) has to be taught to make the same distinction between words having the same word form yet belonging to different word classes. This kind of ambiguity which is known as "homography" (Boot, 7) will be the focus of this paper.

3.1. Words, Sentences, Statements.

In the preceding we made a difference between "words" and "word forms". Both examples, namely the "words" nigger and mean demonstrate that one and the same written unity, called grapheme, are related to different meanings of a single word. If different words are connected with one grapheme the kind of ambiguity is called homography, otherwise the ambiguity is referred to as homonymy.

A sentence in written communication can be defined as a written piece of communication bordered by any punctuation mark not being a comma. In other words the linguistic unity "sentence" can be defined by an external and formal criterion. One does not need any information about the contents of this unity to demarcate it in a text.

A statement, however, can not be so easily located in a text because it is impossible to define this concept without reference to the content of a piece of information.

Example:

"She says that he failed to find a girl friend." It might be evident that the statement of the communicator must be located in the so called subordinated clause "that he failed to find a girl friend".

Example:

"She says that he failed to find a girl friend, and I could observe that she regretted it very much."

For this example the same can be stated as for the first one: statements are: that he failed ...

that she regretted ...

Of course one could imagine more sophisticated examples like:

"She regretted that he failed to find a girl friend."

For this paper, however, the first and second example are sufficient to

allow for the conclusion that the entities "sentence" and "statement" are not equivalent. Apart from this conclusion, a more positive one can also be drawn: A closer examination of both examples indicates that there seems to be an interconnection between the so called "syntactic constituents" of a sentence and the statement. In both cases the statement could be located in the constituent which is called "subordinated clause".

This observation makes it clear that possible statements can be found in locations to be defined in syntactical terms. Besides, these terms are formal rather than content specific. Thus it follows that the objective of this paper can be stated as follows: locate the possible places where statements can be demarcated. A listing of those possible locations would exceed the limits of this paper, our effort is less exhaustive. An impression of those locations, however, is suggested for instance in studying the book by Geoffrey Leech "Semantic and Syntactic Well-formedness". (Leech, 6: 181-185).

Returning to the question of ambiguity, this question can be stated in the following manner: Is it possible to eliminate the ambiguity of words to the extent that the locations of statements in a text can be determined on the basis of available information?

The first question to be answered is: What type of ambiguity will be investigated: homonymy of homography?

Because the statements in the example could both be determined with the help of syntactic information the answer must be that the problem to be solved is homography. Only homography is ambiguity on the syntactic level and should therefore be studied first.

3.2. Homography and Computers.

To date, the problem of ambiguity was analysed in relation to the human brain. Ambiguity proved to be of a rather complex nature. From a systematic point of view, however, ambiguity could be analysed as having a dichotomous structure: homonymy and homography.

As far as the computer is concerned, one has to backtreck one step. For the computer a text is merely a sequence of alphanumeric symbols chained to a string with the length of the whole input text. Each word form can be easily analysed and located in this string. For this purpose one has only to define the alphanumeric symbols not belonging to the word forms

(graphemes) and the alphanumeric symbols not belonging to the graphemes. In other words the computer can be programmed to isolate both graphemes and punctuation marks. Automated content analysis or automatic processing on text level, normally ends after this analysis. The next step consist of the consultation of the dictionary: the graphemes are matched with the entries in the lexicon. More sophisticated procedures provide a step that could be indicated with the linguistic word "lemmatization". In this step, different word forms belonging to one word root are taken together as one word, e.g. mice and mouse are considered to be one word, say MOUSE.

A second more sophisticated step consists of an anlysis of the word types the surrounding words belong to. With this step some homography can be solved and usually the reports on systems working with this strategy are generally very satisfied by the results obtained (kelly, Stone: 8). This strategy could be described as a more linguistically defined way of analysing the problem. It consists of two major steps: the so called tag-declaration and the rules-section (Kelly/Stone,8:84). The ambiguous word forms are provided with all possible "meanings" (= tags). With the help of the rules section one tries to disambiguate the graphemes.

As far as this paper is concerned the most important critique of the strategy is that it only operates on the word level and does not include in fact the unity which we consider as the basic one in communication, namely the statement.

A more general critique of the strategy is that it presupposes that content analysis could be done by context free grammars which strikes us as a contradiction in terms. This type of strategy, however, is very strongly advocated by modern linguistics (Dietrich/Klein, 13), which explains why this linguistic analysis was first applied in content analysis, even though the "pure" linguist does not study problems of text analysis at all. (Boot, 9)

Text analysis implies demarcation of statements. Only in the relevant parts of the text, i.e. the statements to be studied, can the relevant words be found. Only those words should provide the content profile of a text. As far as the computer is concerned there should be devised a step from the grapheme level to the word level. This step should provide the kind of information which allows for the location of the statement

in a text.

What this means for concrete analysis is best answered by an example:

"Chapter 2. The formation of attitudes.

The word attitude has been defined in many ways, none of which, however, differs greatly from what the ordinary individual would understand when he heard or made use of it. An attitude has been defined by Gordon Allport as a mental and neural state of readniness organized through experience, exerting a directive or dynamic influence upon the individual's response to all objects and situations with which it is related." *

The statement we are looking for in this text are:

"as a mental and neural state of readiness organized through experience, exerting .... related"

This statement has the following structure:

1. Mr. Gordon A. says:
   2. mental state of readiness
      2a. modification 1: organized
      2b. modification 2: exerting

In 2b we find further modifications of the word "influence"
the word "response"
the words "objects and situations"

Otherwise stated we first have a broad indication. This indication is modified by so called participle constructions (2a and 2b). In these broader modifications there are further specifications indicated by so called prepositional phrases ("upon ...."; "to ...."; "with ....").

The question we have to answer is therefore: Is it possible to discover automatically this structure of the statement?

Or: what must the computer learn, before it will be possible to discover that structure?

In the words of the previous analysis: the computer should be able to detect participle constructions and prepositional phrases.

The required information can not be derived immediately from a graphematical analysis: after the graphematical analysis of the text only the

---

* Text sample from a large text-corpus for computer investigation in the Netherlands.

graphemes and punctuation marks are known.
As a solution of the problem one could think of a further analysis on the grapheme level. For instance one could analyse word endings. This analysis could be used to find out the places where participles occur: Indeed this kind of analysis proved to be possible and every computational linguist devises his own analyser. As an example, the analyser designed by Terry Winograd will be submitted in appendix 1. (Winograd, 1, 3: 74)
Studying the following utterances, however, one must come to the conclusion that this type of analysis is not enough:

: I am reading a book
: a mental state of readiness exerting a directive influence upon the response

The first participle fullfills a different function from the second one. For the statement we are analysing, only the second type of function, an adjective function is relevant. The function of the participle, however, can only be derived from the fact that in the first statement the participle is preceded by "am" and in the second one the participle is preceded by a noun ("experience").
In other words: it is not enough to know what type of word the grapheme is or may be. The computer must be able to detect the syntactical function of the grapheme.
This observation can be generalized in the following way: an algorithm must be designed with the follwoing input and output specifications:
Input is the output file of the grapheme analyser.
Output is a file containing the actual word classes the graphemes belong to.

3.2.1. Conclusion.

The question about what the computer has to learn, before it will be possible to discover the structure of statements, should be answered as follows:
The computer should learn to assign word classes to graphemes.
As far as homography is concerned no further information either semantical nor syntactical information is needed.

3.3. Assigning word classes by computer.

The assignment of word classes as a field of seperate study is neglected by computational linguistics, because it is assumed to be an impossible task for the computer. My research, however, proved that this assumption is not right and based on a inappropriate design for automated linguistic analysis (Boot, 9).

The model I designed for the automatic assignment of word classes consists of the following steps:

1. The grapheme analysis.
2. The assignment of structure words with the help of a lexicon.
3. The assignment of the remaining syntactical classes with the help of the information provided by 1 and 2. (Boot, 12)
4. The disambiguation of the structure word with the help of 3 (Boot, 9).

3.3.1. Comments: Structure words and content words.

To understand the model for automatic disambiguation of graphemes one should be able to discover the function of the different word classes. The difference between structure words and content words is an old established one in linguistics. This difference is used in automated content analysis and in automated semantic analysis to suppress the structure words. In my proposal these words are given back their vital function in information processing: from these words the place where content words occur are being derived. The ambiguity of the grapheme level, the homography, is solved at the same time as far as the content words are concerned.

Having written a separate paper on the structure word and its function in this type of analysis (Boot, 10), I should like to confine myself here to an enumeration of the word types classifying structure words:

: determiner (the, these, etc.) Code: DET
: preposition (from, of, etc.) Code: PRP
: conjunction (and, as, because, etc.) Code: CON for subordination
and COP for coordination
: pronoun (he, which, none, his, etc.)
Code: PP for personal pronoun
PPI for indefinite pronoun
QD for interrogative pronoun

REL for relative pronoun
POS for possessive pronoun
: auxiliaries (was, should, must, etc.)
Code: VM for auxiliary like should (modularity)
VBC for auxiliaries "have, be"
VP2C for past participle of VBC
VMI for infinitive of VM
VBCI for infinitive of VBC

After the application of the first and the second step in the model, the graphemes as well as the structure words are found and retrieved. For our example the input file for the third step looks like the following diagram:

CHAPTER		DIFFERS	
2	NUM	GREATLY	
	PUNT	FROM	PRP
THE	DET	WHAT	QD
FORMATION		THE	DET
OF	PRP	ORDINARY	
ATTITUDES		INDIVIDUAL	
	PUNT	WOULD	VM
THE	DET	UNDERSTAND	
WORD		WHEN	QD
ATTITUDE		HE	PP
HAS	VBC	HEARD	
BEEN	VP2C	OR	COP
DEFINED		MADE	
IN	PRP	USE	
MANY	PPI	OF	PRP
WAYS		IT	PP
	KOMMA		PUNT
NONE	PPI	AN	DET
OF	PRP	ATTITUDE	
WHICH	QD	HAS	VBC
	KOMMA	BEEN	VP2C
HOWEVER		DEFINED	
	KOMMA	BY	PRP

GORDON	NAAM
ALLPORT	NAAM
AS	VGP
A	DET
MENTAL	
AND	COP
NEURAL	
STATE	
OF	PRP
READINESS	
	KOMMA
ORGANIZED	
THROUGH	PRP
EXPERIENCE	
	KOMMA
EXERTING	
A	DET
DIRECTIVE	
OR	COP
DYNAMIC	
INFLUENCE	
UPON	PRP
THE	DET
INDIVIDUAL	
S	LET
RESPONSE	
TO	PRP
ALL	PPI
OBJECTS	
AND	COP
SITUATIONS	
WITH	PRP
WHICH	QD
IT	PP
IS	VBC
RELATED	

A rough examination of the diagram immediately reveals that we have much information at hand before the third step of the model has to be designed. 1. Of 74 graphemes 32 are structure words. As far as the word classes are concerned they are known.

2. From the information about the punctuation marks we can deduce the type of sentence at hand. From the type of sentence, in turn, considerable information can be derived of the sentence.

Example:

1. He heard the voice of his wife.
2. Do birds fly?

In the first example we know that we are dealing with a so called "declarative sentence". In the second example the punctuation mark indicates that we are dealing with an interrogative sentence.
With respect to the first sentence we know implicitly that we are dealing with the following string structure:

1. noun group followed by
2. verb group

The second example indicated that the first noun group is to be found after the first verb group.
The algorithm that should be designed to assign the word classes for the first sentence must have the following structure:

A. The finite verb.

1. Look for the first noun group (NP).
2. Look for the first open place after this NP, i.e. (OP).
3. Are NP and OP contiguous so OP is the finite verb.

B. The rest of the sentence.

The information at hand contains the following implicit information: The rest of the sentence can only consist of nominal groups or their equivalents. The equivalent of a nominal group is a so called subordinate clause. This information can be derived from the fact that a sentence has only one finite verb. The verb group can consist of a finite verb and one or more NPs.
Inspecting of the sentence reveals only the possibility of NPs, not the possibility of clauses.
The NP has a left hand marker. The left hand marker is a kind of command. It can be translated into the operation: look for the noun.

For once a left hand marker is encountered the noun must follow. In our example no embeddings of NP into NPs are allowed. Therefore, if we encounter the next left NPmarker in the grapheme immediately preceding, this marker (being of course unknown) must be a noun. Up till now we designed the algorithm to the point where "voice" is found. "Voice" will be coded as noun. The only grapheme that is not coded now is the word "wife". This grapheme accurs after the finite verb. In other words it is met in a surrounding where only a noun group is to be expected. The grapheme is met after a double NP left marker: a so called preposition and a possessive pronoun. From that must be concluded that the grapheme "wife" is a noun.
The algorithm that should be designed to assign the word classes for the second utterance must have the following structure:

1. Look for the first verbal group (VB).
2. Look for the first open place after this VB, i.e. (OP).
3. VB and OP are contiguous thus OP is a noun.
4. If the remaining grapheme is the last word of the sentence, than the grapheme is the infinitive form of a verb (VBI).

Because the string properties of sentences are well established, studied profoundly and described by linguistics in the last two thousand years, these are the properties which should be used as a basis to design the algorithm for the assignment of word classes. It is this information which provides the logical shape of the algorithm. To stress again, the logical shape of the algorithm should not merely be based on a so called context free or immediate constituent (IC) grammar (Boot, 11).
Our model is a model for information processing using linguistic knowledge rather than a model for linguistic processing using informational knowledge. We are not engaged in modelling an abstract grammar, but in the simulation of how people proces language materials.

3.3.2. Conclusion.

The preceding chapter strongly suggests that it may be worthwhile to proceed in the proposed modelling fashion. Because linguistics proceeds the other way around and designed first a grammar as a static external instrument, from a linguistic point of view there is no

evidence that it should be impossible to assign word classes in the proposed way. Therefore, the model should be continuously adapted until it can be proven that it does not work.

3.4. Results.

This paper does not provide the linguistic definition of the components of the algorithm because it is not written for computational linguists but for social scientists. Therefore, we demonstrate the results of the algorithm of step 3 of the model with respect to the selected example.

CHAPTER	ADV	WHAT	QD
2	NUM	THE	DET
	PUNT	ORDINARY	ADJ
THE	DET	INDIVIDUAL	SUBST
FORMATION	SUBST	WOULD	VM
OF	PRP	UNDERSTAND	VBI
ATTITUDES	SUBST	WHEN	QD
	PUNT	HE	PP
THE	DET	HEARD	VB
WORD	ADJ	OR	COP
ATTITUDE	SUBST	MADE	VB
HAS	VBC	USE	SUBST
BEEN	VP2C	OF	PRP
DEFINED	VP2	IT	PP
IN	PRP		PUNT
MANY	PPI	AN	DET
WAYS	SUBST	ATTITUDE	SUBST
	KOMMA	HAS	VBC
NONE	PPI	BEEN	VP2C
OF	PRP	DEFINED	VP2
WHICH	QD	BY	PRP
	KOMMA	GORDON	NAAM
HOWEVER	ADV	ALLPORT	NAAM
	KOMMA	AS	VGP
DIFFERS	VB	A	DET
GREATLY	ADV	MENTAL	ADJ
FROM	PRP	AND	COP

NEURAL	ADJ
STATE	SUBST
OF	PRP
READINESS	SUBST
	KOMMA
ORGANIZED	ADJ
THROUGH	PRP
EXPERIENCE	SUBST
	KOMMA
EXERTING	ADJ
A	DET
DIRECTIVE	ADJ
OR	COP
DYNAMIC	ADJ
INFLUENCE	SUBST
UPON	PRP
THE	DET
INDIVIDUAL	SUBST
S	POS
RESPONSE	SUBST
TO	PRP
ALL	PPI
OBJECTS	SUBST
AND	COP
SITUATIONS	SUBST
WITH	PRP
WHICH	QD
IT	PP
IS	VBC
RELATED	VP2

3.4.1. Comments.

The output sample indicates that the algorithm was successful as far as the assignment of word classes to the unknown graphemes (see: diagram 1) is concerned. The algorithm itself is part of a system developed for different european languages (Boot, 12). For that reason some codes have to be explained: the code SUBST is used for NOUN; the code NAAM is used for a proper name.
Table 1 lists the different codes and their meanings

Table 1:

Code	Meaning	Example
SUBST	noun	formation
ADJ	adjective	ordinary
VP2	past participle	defined
ADV	adverb	greatly
VB	finite verb	heard

A closer examination of the output file learns that the encoding is functional indeed. That is the reason why "word" in: "the word attitude" becomes the code ADJ. For the same reason the present participle "exerting" is encoded as ADJ.
For structure words only, one code is changed: the code LET for "s" in "individual's" is replaced by the code POS which is indeed correct.

3.4.2. What information is now available for automated content analysis?

The model we designed consists of 4 steps. Three of them are outlined in the preceding part of this paper. The 4th step remains to be applicated: the disambiguation of the structure word. Indeed there are ambiguities in the structure words as well as in the content word on the word class level. This ambiguity can be established by the wrong encoding for "what" and "when" in:

differs greatly from what the ordinary individual
would understand when he heard or made use of it.

A proper encoding should be: for "what" relative pronoun (REL)
for "when" conjunction (CON)

Before we examine this problem more closely, I wish to investigate the need of information about the structure word for the solution of our problem. It was: delineate the statements in a text! As far as our example is concerned this problem turned out to be equivalent to the

following imperative: demarcate the places where the broader indications and further specifications begin! In linguistic terms this imperative could be translated into: demarcate the places where participle constructions (i.e. modification of broader indications) and prepositional phrases (i.e. further specifications) begin! Since, if the place of modification and specification is discovered the indication (say definition) is implied as well. To demarcate these places one has to know what function the graphemes "organized" and "exerting", i.e. being the participles, have. Apart from this, one should have at one's disposal a grammar for nominal phrases in English.
From the output listing, it can be concluded that all the information we need to reach the statement level of a text is now available: we know the participles and their functions, we also know the sequential positions where prepositional groups begin, and we also know where the nouns belonging to the prepositions are to be found. In other words, we can come to the amazing

3.5. Final conclusion.

An automated content analyser needs only the passes one, two and three of the model.* After the application of these passes, the algorithms which extract the statement can be designed.

---

* The fourth step in the model is the disambiguation of the structure word. As can be concluded from the preceding pages of this paper, this step need not to be described in this paper and I shall confine myself to a list of references (Boot, 9).

4. Summary.

- Content analysis is equivalent to text analysis.
- The basic unit of a text is a statement, not a sentence.
- The problem of ambiguity as far as automated content analysis is concerned has the following structure:
  1. Ambiguity on the grapheme level. The grapheme or word form can belong to different word classes.
  2. Ambiguity on the word level. The word, i.e. the grapheme + the appropriate word class code, may have different meanings.
- The first problem to solve is the problem of ambiguity on the grapheme level: i.e. the problem of homography.
- The problem of ambiguity as far as homography is concerned has the following structure:
  1. Ambiguity in the structure word, the syntactical markers of a sentence. (Assigning word classes to structure words)
  2. Ambiguity in the content word. (Assigning word classes to content words)
- The model proposed in this paper to solve homography consists of the following steps:
  1. Collection of grapheme and other graphematic information, like punctuation marks.
  2. Assigning structure words from a lexicon.
  3. From the information in 1 and 2 and algorithm infer the right word classes of the content words.
  4. From the information in 1, 2 and 3 the ambiguity of the structure word will be solved.
- As far as automated content analysis is concerned the 4th step of the model can be omitted. The information of 1, 2 and 3 together is sufficient to design an algorithm to locate the statements in a text.

References.

1. Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie: The General Inquirer. MIT Press, 1966.

2. David Hays: Introduction to Computational Linguistics. New York, 1967.

3. T. Winograd: "Understanding Natural Language". in: Cognitive Psychology, 1972.

4. Schank, Colby, Weizenbaum: "The Weizenbaum Controversy". in: SIGART, nr. 59, August, 1976.

5. Kallmeyer, Klein: Einführung in die Textlinguistik. Fischer Athenäum, 1976.

6. Geoffrey Leech: Semantics, 1974.

7. M. Boot: "Linguistic Data Structure, Reducing Encoding by Hand and Programming Language". in: Churchhouse, Jones: The Computer in Linguistic and Literary Research, Cardiff, 1977.

8. Edward Kelly, Philip Stone: Computer Recognition of English Word Senses. North Holland Linguistic Series, nr. 13, 1975.

9. M. Boot: Homographie. Ein Beitrag zur automatischen Wortklassenzuweisung in der Computerlinguistik. Diss. Utrecht, 1978.

10. M. Boot: "Key Words in Natural Languages". in: Annals of Systems Research, 1977, pp. 111-121.

11. M. Boot: "PASP: Some Views on Automated Syntactical Parsing of Large Language Corpuses". in: ITL, 23, 1974.

12. M. Boot: "An experimental Design for automated syntactic encoding of Natural Language Texts". in: ALLC-Bulletin, 1977.