GUSTAV LEUNBACH

MORPHOLOGICAL ANALYSIS AS A STEP IN AUTOMATED SYNTACTIC ANALYSIS OF A TEXT

Introduction.

The general purpose of this study is to investigate the possibility of an almost completely automated syntactic analysis of a given text of a known language. This has in itself some theoretical linguistic interest, and to the extent that it succeeds, it will save a large amount of labour in relation to automated indexing and translation, etc..

One step in this analysis is to count the occurrences of all word-forms of the text, not for statistical purposes, but for the purpose of selecting a list of highly frequent word-forms for which the labour of adding syntactic information before the automated analysis is most fruitful.

This list contains some substance words for which the syntactic information mainly consists of assigning a word class; this helps the automated analysis just because the words are frequent; they appear in many periods of the text.

But mostly the list contains structural words (prepositions, pronouns, adverbs, auxiliary verbs, etc.); without information on the syntactic function of these words it is usually impossible to analyse a text at all.

Another step in the investigation – the one with which this paper deals – is the automated utilization of the morphological information contained in the text: a search of the implicit information on word class and syntactic function which is given by the knowledge that two or more word-forms are flexions of a common stem.

The concept of a lemma.

As indicated above, a lemma is defined as a group of two or more word-forms of the text which morphologically may be flexed forms of a common word stem within one of the flexion paradigms included

in the analysis (most so-called irregular paradigms being excluded as too difficult to automate).

The language in question is Danish, and a list of 25 paradigms and paradigm variants for that language is given as Table 1.

1	E	ES	–	S	ET	ETS	ENE	ENES						
2	E	ES	–	S	EN	ENS	ENE	ENES						
3	E	ES	Z – Z	S	ET	ETS	ENE	ENES						
4	E	ES	Z – Z	S	EN	ENS	ENE	ENES						
5	E	ES	X – X	S	ET	ETS	ENE	ENES						
6	E	ES	X – X	S	EN	ENS	ENE	ENES						
7	ER	ERS	–	S	ET	ETS	ERNE	ERNES						
8	ER	ERS	–	S	EN	ENS	ERNE	ERNES						
9	ER	ERS	Z – Z	S	ET	ETS	ERNE	ERNES						
10	ER	ERS	Z – Z	S	EN	ENS	ERNE	ERNES						
11	ER	ERS	X – X	S	ET	ETS	ERNE	ERNES						
12	ER	ERS	X – X	S	EN	ENS	ERNE	ERNES						
13	E	ES	–	S	ET	ETS	NE	NES						
14	E	ES	–	S	EN	ENS	NE	NES						
15	E	ES	ER	ERS	ET	ETS	ERNE	ERNES						
16	E	ES	ER	ERS	EN	ENS	ERNE	ERNES						
17	EDE	EDES	–	S	ET	ETS	ENDE	ENDES	R					
18	E	ES	ETE	ETES	ET	ETS	ER							
19	E	ES	EDE	EDES	ET	ETS	ENDE	ENDES	ER	–				
20	E	ES	EDE	EDES	ET	ETS	ENDE	ENDES	ER	X –				
21	E	ES	T	TS	TE	TES	ENDE	ENDES	ER	–				
22	E	ES	X T	X TS	X TE	X TES	ENDE	ENDES	ER	X –				
23	E	ES	ST	EST	ERE	ERES	ESTE	ESTES	T	–	STE	STES		
24	E	ES	ST	EST	ERE	ERES	ESTE	ESTES	X T	X –				
25	E	ES	ST	EST	ERE	ERES	ESTE	ESTES	Z T	Z –				

TABLE 1

25 PARADIGMS AND PARADIGM VARIANTS OF STANDARD DANISH

THE FIRST 16 ARE NOMINAL PARADIGMS, NOS. 17 TO 22 ARE VERBAL, THE LAST THREE ADJECTIVAL. THE ORDER OF ENDINGS WITHIN EACH LINE HAS NO SIGNIFICANCE WHATEVER

X AND Z BEFORE ENDINGS INDICATE STEM MODIFICATIONS
X THE STEM IS FORMED BY DOUBLING THE LAST LETTER BEFORE THE ENDING
Z AN E BEFORE THE LAST LETTER BEFORE THE ENDING (FLUID E) IS DELETED AND A POSSIBLE DOUBLE LETTER BEFORE THIS E IS DISSOLVED.

The text has been read into a computer, giving each word-form an index number the first time it occurred and storing in one file the text represented by index numbers and in another the word-forms. Lemmatization was performed on this latter file by reading each word

backwards. If the last letter was one which could form an ending or a part of an ending the preceding letter was also searched, etc..

All cases of a possible stem and ending have been treated as putative stems. To quote the extreme example, a form XXSTES is treated as a putative stem in itself and as derived from the following putative stems: XXSTESS (this is the *x*-modification of Table 1, the double consonant only appears before endings containing an *e*); XXSTS (the *z*-modification); XXST; XXS and XX. The last three all occur in common Danish words; the three first do not, but this only means that another derivation from the same putative stem does not appear unless by a very rare coincidence.

The putative stems are stored in a list, and each one is compared with all preceding until it is revealed as either one of them or a new one. It is essential for the practical maintenance of the analysis that the word-forms are sorted at least by first letter and the comparison made one group at a time; actually the words were ordered nearly alphabetically, but no general rule could be devised for stopping the search earlier.

After the search has been completed, all putative stems which appear only once are discarded. Of the rest, only those are conserved where at least two endings belong to the same paradigm, that is to the same one among the 25 lines of Table 1. Finally, a file of the index numbers of all word-forms is loaded with information on every lemma that each word-form belongs to and with which ending. If the same word-form belongs to more than one lemma, there is a conflict, see later.

The text is a short Western novel of about 41,000 words which have been sorted into slightly more than 5,000 word-forms. Of these, about 2,950 have been marked as members of lemmas.

However, this does not mean that useful syntactic information can be derived for nearly 60 pct. of the word-forms of the text. Unless the lemma is a pure coincidence (homography between two unrelated words apart from some letters which may be flexion endings; this is a problem of the same kind as ordinary homography is for any automated text analysis), it indicates a semantic bond between its members, but the syntactic information depends on the set of endings contained in the lemma.

In the following, lemmas will be divided into three main categories, unambiguous, ambiguous and conflicting. So far, a complete analysis has been performed on the 323 lemmas formed by words with initial letters A-G, including 882 out of 1568 word-forms.

Unambiguous lemmas.

A lemma is called unambiguous if the endings of its members determine unambiguously which of the three word classes, noun, adjective and verb, the lemma may belong to. Normally, this also means that the endings can only belong to one of the 25 paradigms of Table 1.

Of the 323 lemmas formed with the initials A-G, 162 are classified as unambiguous.

Nouns account for 79 lemmas. Most of these contain the ending *-en*, the definitite article in singular for nouns of the common gender, which does not belong to any adjectival or verbal paradigm; several contain the plural definitite article.

One lemma has 6 members; a computer-aided inspection has revealed this to be the word "bandit", a very appropriate word for this type of text, several of its forms have high frequency. Two lemmas have each 5 members and seven have 4 each, but more than half have only 2 members. With only two forms represented, the risk of a chance homography is of course high.

Adjectives account for 14 cases; they nearly all include the comparative ending *-ere* or the superlative *-st* which can be extended with *-e* or *-es*. A 15th lemma was also by the computer classified as unambiguously adjectival, but it contained only two forms with the endings *-ste* and *-stes*, and as the example above of different borders between putative stem and ending shows, either *-s-* or *-st-* may belong to the stem. (Inspection showed the word actually to be a verb with a stem ending in *-s*.)

Two of the fourteen lemmas have 4 members each and eight have 3.

Verbs: 69 lemmas. 22 of these include the ending *-ø* which in almost all verbal paradigms has the function of imperative which is unlikely to occur so frequently. Instead, the form with zero ending is probably in most cases a noun of the same stem, but not represented in any of those forms which differ from verbal flexion endings, and thus not causing a conflict between two lemmas.

Two verbal lemmas have 6 members each (both including the dubious imperative), one has 5 and ten have 4 (six of them including the zero form, the dubious imperative); almost half have only 2 members.

If a lemma is by this definition unambiguous, and the set of endings contained in it gives no reason to suspect its validity, the syntactic information on each word-form is in most cases quite obvious. With

nouns and adjectives, the forms ending in *-s* are of particular interest; they are possessive forms with function of adjectives, but with a peculiar distribution of articles. In verbs the participles may form composite tenses (the present participle almost never) or may function as adjectives with restricted flexion patterns.

Ambiguous lemmas.

These 115 lemmas have none of the forms which are characteristic of either of the three word classes; all are short, 13 of them have 3 members and the rest only 2.

58 are classified as possibly nouns or verbs, but 37 of these contain the zero ending, mostly together with *-et* which may be either the definite article of a neuter noun or a past participle, or with *-er* which may be plural or present tense. As the flexion-less form of a noun is much more frequent than the imperative of a verb, the large majority of these words are certainly nouns. It may be reasonable to classify them together with some of the shorter nominal or verbal lemmas as "almost unambiguous".

Conflicting lemmas.

The same word-form may be a member of more than one lemma, either with the same border between stem and ending or with a different one. In each of these cases, each lemma is followed through until all word-forms interconnected directly or indirectly are summed up, and the whole complex is defined as one conflicting lemma.

There are 46 conflicting lemmas in the part of the vocabulary investigated, containing from 3 to 7 members, apart from one of 10 members which on inspection proved to be a mixture of demonstrative pronouns.

Most of the others can be explained either by a noun and a verb formed from the same root (occasionally also an adjective) or by homography between two unrelated words, each with some flexed forms represented.

When both one or more of the shorter endings of a paradigm and one of the longer endings are present, one may with reasonable certainty assess the word class and syntactic function of the longer form,

even if the shorter form may belong to more than one paradigm and no inference can be drawn on its word class.

Danish compared to other languages.

The same analysis performed on a text in English will be somewhat easier in computational work because of the fewer paradigms and fewer forms in each, but I expect the yield to be much less. I would expect a lot of ambiguous lemmas (noun or verb), containing the zero form and the ending *-s*, some verbal lemmas, containing *-ed* or *-ing* or both, and a few adjectival lemmas with comparative or superlative endings. A noun can only be identified by the endings *-'s* or *-s'*, which I expect to be rare in most texts – and which can be found by more mechanical means also. Many verbal lemmas would be expected to hide a noun of the same root.

On the other hand, applied to a language with richer flexions the computational work involved in this analysis might get out of hand, and I am not sure that the yield would in all cases increase comparably.

In German I expect many ambiguous lemmas to be found because several endings occur in paradigms of different classes. French has the same problem as English that nouns can hardly be identified by their endings; moreover the boundary between regular and irregular paradigms is rather vague in written French, making it somewhat uncertain how many paradigms it pays to include in the automated analysis.

Russian seems to me to offer somewhat better possibilities. There are many verbal paradigms with various stem modifications which it may be impossible to automate, but the forms of the present tense on one hand, and most other forms separately may make it possible to identify a verb by two different stems. The risk of finding ambiguous lemmas is small, though conflicting lemmas may occur. And the specific information on case which many nominal forms contain is often relevant to the syntactis analysis, although it may be difficult to utilize it in an automated manner.

S. Hellberg's analysis of Swedish.

After the computational work for this paper had been completed my attention was drawn to an article by S. HELLBERG: *Computerized*

Lemmatization without the Use of a Dictionary, in « Computers and the Humanities », VI (1972) 4, pp. 209-212. The following discussion of his method is built mainly on a stencilled report in Swedish mentioned in a footnote to the article.

Hellberg does not investigate all possible ways of dividing one word-form into stem and ending, but investigates for two word-forms at a time whether they can be the same stem with two flexion endings within the same paradigm.

This procedure depends heavily on a complete alphabetization of all word-forms. In the cases where Swedish has the same stem modifications that my program takes into account, Hellberg has to define the stem as the shortest part of the word which is identical in all flexed forms; this increases the number of different endings, and the definition of a stem comes rather far from the ordinary grammatical concept, while mine corresponds rather closely to it.

Hellberg's method may be faster than mine in computation time, but I believe that he runs a greater risk of missing lemmas which are actually present than I; the concept of a conflicting lemma does not appear in his system, instead he discusses the frequency of wrong lemmatizations.

Practical results.

Of a vocabulary of slightly above 5,000 word-forms in a novel of about 41,000 words, nearly 60 pct have proved to be members of lemmas, that is sets of two or more word-forms which may be flexions of the same stem within one of the regular paradigms.

An inspection of those lemmas which are formed by words with initials A-G shows that of the word-forms contained in them, about 60 Pct can with reasonable certainty be assigned to a definite word class, about half as nouns, somewhat less as verbs, and a few as adjectives. The rest cannot be thus assigned, either because all the endings of the lemma may belong to paradigms of different word classes, or because the same word-form belongs to lemmas of different paradigms, thus indicating the possibility of a homograph.

Of the endings pertaining to a given paradigm, some convey a definitite syntactic function while others do not.

These figures summarize the yield of the analysis as regards its function as a step in an automated syntactic analysis of the text.

A different summarization of the results of the analysis, which has some general interest, concerns the distribution of lemmas by length. The part of the lemmas investigated in detail show three lemmas of 6 members and three of 5 members as the longest.

Which words occur in so many flexions depends of course on the content of the text, but the numerical structure I suppose to be characteristic of a much wider range of self-contained texts; if the investigation is carried out on the vocabularies of different texts, an interesting by-product may be the study of the stability of this structure.

One may be interested in a different type of statistical analysis, namely of the frequencies of various grammatical categories, for instance, the frequencies of singular and plural forms of nouns, of present and past tense. For this purpose, however, the analysis performed here is not sufficient, not even combined with the frequency count of all word-forms mentioned in the introduction: firstly, the analysis expressly only recognizes a flexion ending if at least two different endings to the same stem occur; that is, the frequency of e.g. plural forms may be counted only for those nouns which are represented in at least two forms. Secondly, in the Danish nominal paradigms not all forms are with certainty identifiable as either singular or plural, and in the regular verbal paradigms the simple past is always the same form as a participle which is used in an adjectival function, and not in composite tenses.

(For the purpose of test parsing, the latter ambiguity is not a fundamental obstacle; whenever such a form is encountered two possibilities must be left open, and one must hope that other words of the period will close one possibility).