HEINZ J. WEBER

THE AUTOMATICALLY BUILT UP HOMOGRAPH DICTIONARY - A COMPONENT OF A DYNAMIC LEXICAL SYSTEM -

O. *Introduction.*

Ambiguous word forms (often called " homonyms " or – in written language – " homographs ") are known as obstacles in many fields of computational linguistics, especially in automatic documentation, content analysis or mechanical translation. In this respect two problems must be distinguished:

1) the detection of homographic word forms in the text,
2) their disambiguation by analysis procedures.

This paper exclusively deals with the first problem.

1. *The Detection Of Homographs.*

1.1. In current procedures for the detection of homographs two alternatives can be differentiated:

i) Homographs are identified like monosemic word forms by segmentation and looking up in the standard lexicon. Homographs are detected, if segments of text word forms correspond with more than one lexicon-entry. Lexicon-entries representing homographic items therefore need no special marking.

ii) Homographs are identified by means of a special homograph dictionary, which can be worked out in two versions:

1) the homograph dictionary contains the graphemic shapes of all homographic word forms (full forms) and their possible linguistic specifications. In this case no segmentation procedures are required.

2) the dictionary does not contain full forms but only the respective canonical forms. A special marking gives information about other corresponding dictionary-entries and the extent of their overlapping.

In both cases (1) and (2) the identification of homographic text word forms is separated from the identification of monosemic ones.

Procedures (i) and (ii) have some characteristic advantages and defects, which I will consider rather briefly.

1.2. As already pointed out the first method requires (1) a segmentational algorithm, with the help of which the word forms of a text can be parsed into segments (e.g. stems and inflectional affixes), (2) an identificational component composed of a grapheme-sequence-comparing algorithm and the standard lexicon; thus it can be checked, whether a text segment detected by (1) is the expression-side of one (or perhaps more) lexical unit(s). If this is the case, the content-side(s) of the corresponding unit(s) can be assigned to the text segment.

1.2.1. According to this conception the identification of word forms would offer no problems in the following cases:

i) the word form represents only *one* sequence of segments (is monosemic), that means that each segment corresponds to *one* lexicon-entry:

Germ.:	...ØkindesØ...	–/kind/+/es/	*Kind*
Engl.:	...ØchildsØ...	–/child/+/s/	*child*
Frnc.:	...ØenfantØ...	–/enfant/+/Ø/	*enfant*
Russ.:	...ØrebënokØ...	–/rebënok/+/Ø/	*rebënok*

ii) The word form can be parsed into more than one set of segments (is homographic), and the possible readings show coinciding segment-boundaries:

Germ.:	...ØlautØ...	–/laut/+/Ø/	*Laut* (*SUB*)
		/laut/+/Ø/	*laut* (*ADJ*)
		...	...
Engl.:	...ØmeanØ...	–/mean/+/Ø/	*mean* (*ADJ*)
		/mean/+Ø/	*to mean*
		...	...
Frnc.:	...ØmortØ...	–/mort/+/Ø/	*mort* (*SUB*)
		/mort/+/Ø/	*mort* (*ADJ*)
		...	...
Russ.:	...ØžilaØ...	–/žil/+/a/	$žila_1$
		/žil/+/a/	$žila_2$

If the detected segments are compared with all graphematically corresponding lexicon-entries, that is, if the lexicon-look-up is not stopped after the first correspondence, this case can also easily be coped with. If the entries are arranged in alphabetical order, the respective units will immediately succeed one another.

1.2.2. The detection becomes difficult however, if one word form is parsed into more than one set of segments (is homographic), while the segment-boundaries in the possible readings are overlapping:

Germ.:	...ØgetriebenØ...	-/trieb/+/ge-en/	*treiben*
		/getriebe/+/n/	*Getriebe*
Engl.:	...ØhearingØ...	-/hear/+/ing/	*to hear*
		-/hearing/+/Ø/	*hearing*
Frnc.:	...ØpêcherØ...	-/pêch/+/er	*pêcher (VRB)*
		/pêcher/+/Ø/	*pêcher (SUB)*
Russ.:	...ØvalaxØ...	-/val/+/ax/	*val*
		/valax/+/Ø/	*valax*

As the segments which have thus been detected do not coincide graphematically (e.g. /trieb/ – /getriebe/), i.e., as the respective lexicon-entries are to be found at different places in the lexicon, for the identification of such homographs enormous parsing – and comparing – procedures are required. As cases of homography with overlapping segment-boundaries in the various readings are encountered quite frequently in languages with extensive inflection (e.g. German, French, Russian), method 1.1. (i) is not the best in any case.

1.2.3. The advantage of this method is above all to be seen in the fact that the identification of homographs can be managed automatically, and that no special marking of the respective entries is necessary. This is especially important with regard to dynamic lexical systems, where the number of lexicon-entries and their specification can vary; new entries do not require a change of the detection procedure. The disadvantage consists in the fact that monosemic and ambiguous word forms are submitted to the same procedure, which amounts to an undue delay of the determination of monosemic word forms. Multiple parsing with subsequent lexicon-look-up has always to be applied if the respective text word forms contain grapheme sequences, which correspond to inflectional affixes. Only after this can it be found out whether more than one plausible reading has resulted:

Germ.:	...ØgetriebenØ...	-/trieb/+/ge-en/	*treiben*
		/getriebe/+/n/	*Getriebe*

or only one reading is true:

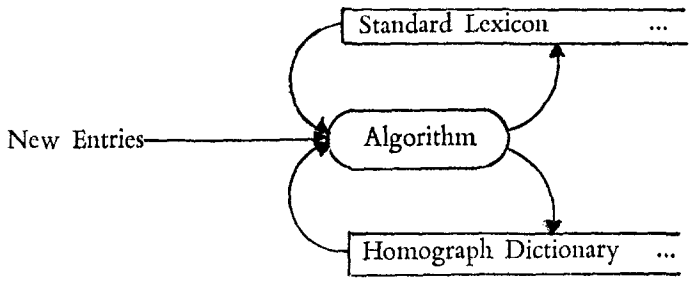
Germ.:	...ØgelagenØ...	-/gelage/+/n/	*Gelage*
		but not:	
		/lag/+/ge-en/	*liegen*

1.3. Method 1.1. (ii) does not have this disadvantage. Homographs are separately registered and marked according to their readings; thus a considerable acceleration of the identificational procedure is made possible. *Monosemic* word forms, the stems of which are listed up in the standard lexicon, are identified more easily, as the segmentation and lexicon-look-up can already be stopped after assignment of *one* reading. *Ambiguous* word forms are specified more easily, as the extensive segmentation- and comparing-procedures do not have to be applied (as the various readings are registered in the homograph dictionary – version 1.1. (ii) (1) –) or are reduced to a minimum (as the respective lexicon-entries bear a special marking, by which their homography can be derived – version (ii) (2) –). These advantages however entail certain disadvantages: as a rule homograph dictionaries are built up manually and have to be manually complemented, when the standard lexicon is extended; the same has to be stated for the marking of lexicon-entries. Aside from this troublesome and time-consuming business one cannot be sure that all homographies are registered or are marked exhaustively.

2. *The Automatically Built Up Homograph Dictionary.*

2.1. In this paper a method will be outlined, in which the advantages of the first procedure are combined with those of the second one: the standard lexicon therefore can be extended automatically without delaying the identificational procedure. The homograph dictionary is compiled by analysis of the standard lexicon; all stems representing homographic items are taken away from it and integrated into the homograph dictionary. The same algorithm, which detects homographies incorporated in the standard lexicon, can be used to find out by analysis of both lexica, whether new entries and all inflected forms represented by them are homographs. If this is the case, they are registered in the homograph dictionary, otherwise in the standard lexicon. Thus the number of entries in both lexica can be increased

automatically and the specifications of ambiguities in the homograph dictionary always correspond to the current state of information.



2.2. The standard lexicon can be characterized as follows: the lexicon-entries are "stems"; each stem representing a set of inflected forms, which is called a "paradigm" or "part of a paradigm". In order to abbreviate the graphemic assimilation of text word forms to the graphemic shapes of lexicon-entries during the identificational procedure, morphologically and syntactically determinated allomorphs have also been noted. Stems of complex lexical units (e.g. /kickØtheØ bucket/, /zumØzugØkomm/) are, however, ignored:

stems of (Ger.)	*graben* (*VRB*):	/grab/, /grub/
	beipflichten:	/beipflicht/, /pflicht/ ...
stems of (Eng.)	*to sing*:	/sing/, /sang/, /sung/
stems of (Frc.)	*mourir*:	/mour/, /meur/, /mort/, ...
stems of (Rus.)	*rebënok*:	/rebënok/, /rebënk/, ...

2.3. The homograph dictionary is built up by comparing selected entries of the standard lexicon. In order to elucidate the comparing procedure we restrict ourselves to the coordination of just *two* lexicon-entries. Two stems represent homographic inflected forms, if the following conditions are fulfilled:

i) The graphemic shapes of the stems belonging to the paradigms P_1 and P_2 are identical:

$$G_s(P_1) \mathrel{\widehat{=}} G_s(P_2).$$

In this case homography exists, if any inflectional affixes co-occurring with the respective stems are homographic too:

$$G_f^i(P_1) \mathrel{\widehat{=}} G_f^i(P_2).$$

ii) The graphemic shape belonging to the stem of paradigm P_1, concatenated with a sequence G_k, is homographic with the stem of paradigm P_2:

$$G_s(P_1) \times G_k \mathrel{\widehat{=}} G_s(P_2).$$

In this case the graphemic shapes of the co-occurring inflectional affixes have to correspond in the following way:

$$G_f^i(P_1) \setminus G_k \mathrel{\widehat{=}} G_f^i(P_2).$$

Concerning G_k some restrictions have to be observed:

iii) G_k has to be homographic or partially homographic with any inflectional affix of the respective language:
$G_k \subseteq G_f^i$, for G_f^i as an element of the finite set $\mathscr{F}$, which contains all inflectional affixes.

iv) G_f^i for its part must be co-occurring with the respective stem of the paradigm:

$$G_f^i \equiv G_f^i(P_1).$$

The co-occurence of stem and affix is specified by the respective inflection-class-marking of the lexicon-entry.

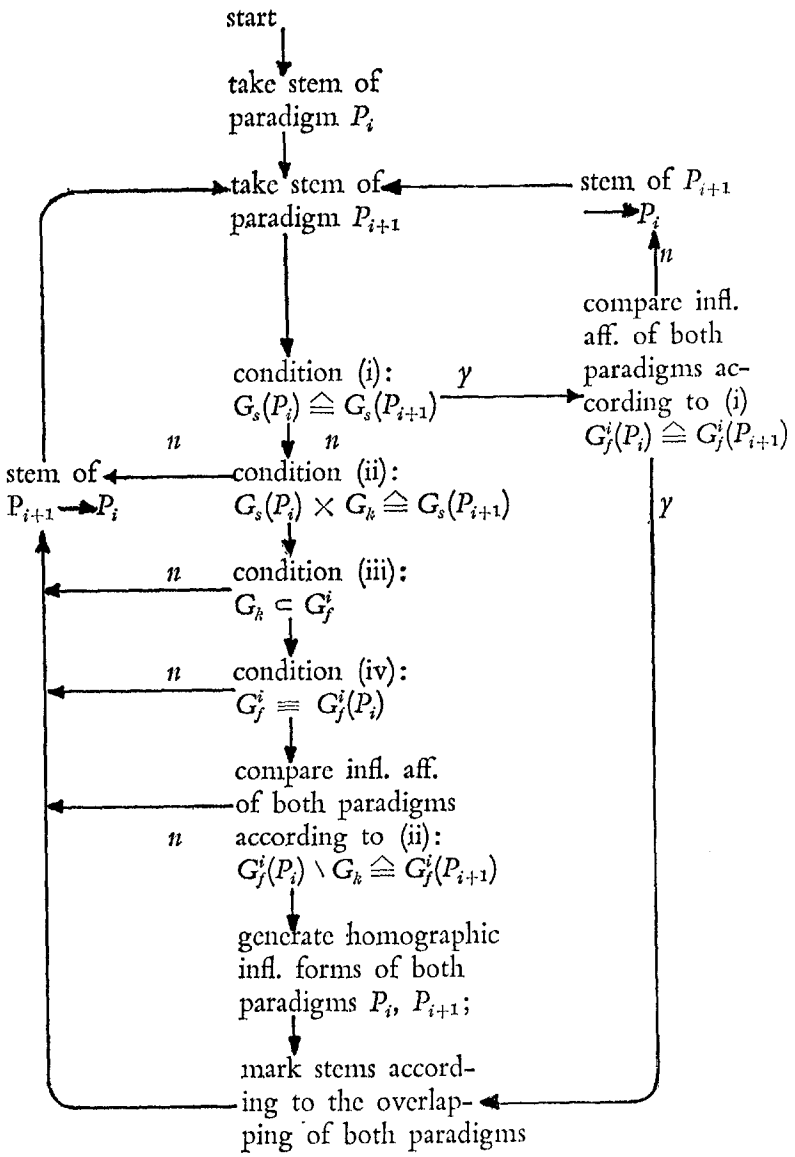
2.4. Conditions (i) and (ii) determine the selection of stems. (iii) prevents that all partially homographic stems in the lexicon are selected and examined. Thus – for example – a coordination of the German stems /arm/ *Arm* and /armee/ *Armee* is not permitted (though they are fulfilling condition (ii): $G_s(P_1) \times$ /...ee/ $\mathrel{\widehat{=}} G_s(P_2)$), as the sequence $G_k =$ /...ee/ does not fulfill the condition $G_k \subseteq G_f^i$.

iv) further reduces the number of the selected stems to those cases, where G_f^i ($\supseteq G_k$) is co-occurring with $G_s(P_1)$ (e.g. German /schwer/ *schwer* and /schwert/ *Schwert* are not combined, though both stems fulfill conditions (ii) and G_k (= /...t/) fulfills condition (iii); but G_k does not fulfill (iv): /...t/ is not homographic with an affix co-occurring with /schwer/ (*ADJ*)).

The relationship between the graphemic shapes of the affixes co-occurring with the stems of both paradigms is finally specified by the complementary part of condition (ii): $G_f^i(P_1) \setminus G_k \mathrel{\widehat{=}} G_f^i(P_2)$. E.g. the German stems /hoer/ *hören* and /gehoer/ *Gehör* (*SUB*) could be combined according to condition (ii): /hoer/ $\times$ /ge.../ $\mathrel{\widehat{=}}$ /gehoer/. Moreover conditions (iii) and (iv) concerning $G_k =$ /ge.../ would be fulfilled: /ge.../ $\subset G_f^i$ and G_f^i(= /ge–t/) $\equiv G_f^i(P_1)$. But there is no inflectional

affix /t/ co-occurring with the stem /gehoer/ of paradigm P_2, which corresponds to /ge-t/ in the mentioned way: /ge-t/ \ /ge.../ ≙ /t/.

2.5. These conditions determining the selection of lexicon-entries and their examination for homographic items can be transferred into programmed instructions:



2.6. Prerequisites for the outlined algorithm are:

1) a computerized stem-lexicon; the entries, which have to be arranged in alphabetical order, must bear an inflection-class-marking, which makes it possible to generate all inflected forms represented by the respective stems:

...
/ficht/ : *fechten, VRB,* infl.–class 48
/fichte/ : *Fichte, SUB,* infl.–class 5
...

2) A complete list of inflectional affixes, which bear the possible inflection-class-markings corresponding to those of the stems:

/Ø/ : *SUB,* infl.–classes ..., 5, ...
VRB, infl.–classes ..., 48, ...
...
/st/ : *VRB,* infl.-classes ..., 48, ...
...

2.7. The selection of stems and the comparison of the co-occurring inflectional affixes could be carried out in a slightly modified way. As already pointed out, the selection of stems is in the main determined by the grapheme sequence G_k (which specifies the graphematic overlapping of non-homographic stems). Further restrictions concern the correspondence between G_k and the inflectional affixes co-occurring with the selected stems (see 2.3. (iii) and (iv). As the inflection-class-markings of stems and affixes (which are similar) are shortened distributional classifications, it is obvious to bring them into a system, according to the respective specifications of G_k. A matrix is built up by which it can be seen whether a G_k -specification restricts the coordination of stems with certain inflection-classmarkings. In this way the detailed examination and comparison of all co-occurring affixes (in accordance to condition 2.3. (iv)) can be substituted by one single operation, at least in a good number of cases.

coordinated infl.-classes of $stem_1/stem_2$	G_k-specifications of $stem_1$: $G_k = /...e/$	...
...	...	...
VRB 1 /*SUB* 5	+	
VRB 48/*SUB* 5	—	...
VRB 55/*SUB* 5	+	
...	...	

SUB 5: /stelle/ *Stelle*; /fichte/ *Fichte*;
/wiese/ *Wiese*; ...
VRB 1: /stell/ *stellen*; ...
VRB 48: /ficht/ *fechten*; ...
VRB 55: /wies/ *weisen*; ...

This matrix forbids the coordination of two stems belonging to the classes *VRB* 48 and *SUB* 5 (in this succession), though they may correspond in accordance to condition (ii):

$$/\text{ficht}/ \times /...e/ \mathrel{\widehat{=}} /\text{fichte}/.$$

On the other hand the coordination of two stems belonging to the classes *VRB* 1 or *VRB* 55 and *SUB* 5 is admissibile:

$$/\text{stell}/ \times /...e/ \mathrel{\widehat{=}} /\text{stelle}/ \text{ or}$$
$$/\text{wies}/ \times /...e/ \mathrel{\widehat{=}} /\text{wiese}/.$$

The building-up of such matrices seems to be a useful device, as the number of G_k-specifications in the respective languages (German, French, Russian) is limited. In German we have found out ten frequent and about thirty extremely rare G_k-specifications. In English homographies with graphemetically overlapping stems are without that rather seldom.

2.8. The conceived algorithm selects just *two* entries (respectively their paradigms), which are examined for homographic word forms. After the first cycle of selecting and comparing – as pointed out in 2.5. – homographs, which are members of more than *two* paradigms (e.g.

German ØalbenØ: *Album*, *Alba*, *Albe*, Alb_1, Alb_2), therefore, are not specified exhaustively:

$$\begin{aligned}
Album \cap Alba &= \{\text{alben}\}\\
Album \cap Albe &= \{\text{alben}\}\\
Album \cap Alb_1 &= \{\text{alben}\}\\
Album \cap Alb_2 &= \{\text{alben}\}\\
Alba \cap Albe &= \{\text{alben}\}\\
Alba \cap Alb_1 &= \{\text{alben}\}\\
Alba \cap Alb_2 &= \{\text{alben}\}\\
Albe \cap Alb_1 &= \{\text{albe, alben}\}\\
Albe \cap Alb_2 &= \{\text{alben}\}\\
Alb_1 \cap Alb_2 &= \{\text{alb, alben}\}
\end{aligned}$$

In this example ØalbenØ is described as a member of altogether ten intersection sets (which are the results of the first coordination-cycle). In a second cycle (which will not be dealt with in detail) all intersection sets of the first cycle are examined for identical word forms. ØalbenØ now can be described as an intersection set of five paradigms:

$$\begin{aligned}
Album \cap Alba \cap Albe \cap Alb_1 \cap Alb_2 &= \{\text{alben}\}\\
Albe \cap Alb_1 &= \{\text{albe}\}\\
Alb_1 \cap Alb_2 &= \{\text{alb}\}
\end{aligned}$$

The coordination of intersection sets in the second cycle is – as well as the coordination of paradigms in the first one – determined by conditions, which are derived from the graphemic shapes of the respective stems. In all probability stems like /album/, /alba/, /albe/, /alb/ will represent at least one homographic word form, while stems like /album/, /alge/, /alibi/, /altar/ will not.

2.9. As the outlined method of building up a homograph dictionary is, in the main, using facts of the expression-side of lexical units, it can be applied to various lexicon-types. The content-sides of the respective entries (i.e. stems) can bear either semantically, syntactically or otherwise relevant information.

REFERENCES

H. EGGERS, *et al.*, *Elektronische Syntaxanalyse der deutschen Gegenwartssprache*, Tübingen 1969.

D. KRALLMANN, *Linguistische Datenbank und kumulatives Wörterbuch*, in *Kolloquium Maschinelle Sprachverarbeitung*, Mannheim 1968.

W. LENDERS, *Static And Dymanic Lexical Systems*, Bonn 1969.

H. J. WEBER, *Mehrdeutige Wortformen - grammatische Beschreibung und lexikographische Erfassung* (thesis), Saarbrücken 1973.