WERNER BRECHT

MORPHOLOGICAL ANALYSIS (A FORMAL APPROACH)

1. MOTIVATION AND BASIC IDEAS

Since 1972 a researcher team in Bonn has been working on the automatic syntax-analysis of the german language. The morphological analysis is a part of this work that has already been formalized and programmed by the author. We can consider the following paper as a generalization of this formalized description.

In this first chapter some expressions like " text " or " lexicon " are considered clear by intuition. Later on we'll get to know the exact definitions. The basis of each description of a text of any language is a morphological analysis of this text. One can easily agree that such a description has to be derived from the words or sentences of the text which is to be described. The expression " description of a text " is understood in a very general sense. One can imagine a syntactic description or a semantic interpretation or a combination of both of them or any other information.

In a natural language the number of the possible texts is not finite. That's easy to prove because of the following sentences:

One is a number
Two is a number
...

Hence it's impossible in practice to use a lexicon of the following form

text 1	*description of text* 1
text 2	*description of text* 2
...	...

for all texts of a language. There remains only one possibility. One has to ascribe the words and sequences of words which every text consists of with one or more (homography) descriptions. Then one can try to derive the descriptions of the text out of the descriptions of the words or sequences of words.

2. BASIC DEFINITIONS

2.1. *Remark.*

Let A and B be sets. Then denotes:

$P(A)$: the set of all subsets of A (the powerset of A)
$A^\star$: the free monoid over A
$A \times B$: the cartesian product of A and B
$|x|$: the " length " of $x \in A^\star$ (the number of elements of A which x consists of); $|x| \in IN$.

Now we'll define the expressions " character " and " string " for a language. We use five basic sets:

$LETT$ $:= \{A, B, C, ..., Z\}$ the set of " letters "
DIG $:= \{0, 1, 2, ..., 9\}$ the set of " digits "
$BLANK$: $= \{\llcorner\lrcorner\} = : \{blank\}$, $|\llcorner\lrcorner| = 1$, $\llcorner\lrcorner \in BLANK^\star$
PS $:= \{.\} \cup \{,\} \cup \{!\} \cup ...$ the set of " punctuation-signs "
SS $:= \{\%, \&, \S, ...\}$ the set of " special-signs ".

From these five sets we derive:

a) $CHAR := LETT \cup DIG \cup BLANK \cup PS \cup SS$ the set of " characters ".

If $x \in CHAR$, we say: " x is a character ".

b) $CHAR^\star$: the free monoid over $CHAR$.

If $x \in CHAR^\star$, we say: " x is a string "
or: " x is a sequence of characters "
or: " x is a text "

c) $LISS := LETT \cup DIG \cup SS$ the set of " characters without blank and punctuation-signs ".

d) $LISS^\star$: the free monoid over $LISS$.

If $x \in LISS^\star$, we say: " x is a string without blank and punctuation-signs ".

2.2. *Remark.*

$[LISS \subset CHAR \Rightarrow LISS^{\star} \subset CHAR^{\star}] \Rightarrow [e \in LISS^{\star}, e \text{ empty element} \Rightarrow e \in CHAR^{\star}, e \text{ empty element}]$

Now we define the expression " word ".

2.3. *Definition.*

$WORD1 := \{x \mid x \in BLANK^{\star} \wedge |x| > 0\}$
$WORD2 := \{x \mid x \in LISS^{\star} \wedge |x| > 0\}$
$WORD := WORD1 \cup WORD2 \cup PS$
If $x \in WORD$, we say: " x is a word ".

2.4. *Examples.*

$\sqcup\sqcup\sqcup \in WORD$, because $\sqcup\sqcup\sqcup \in WORD1$, $|\sqcup\sqcup\sqcup| = 3$
WHEN $\in WORD$, because *WHEN* $\in WORD2$, $|$ *WHEN* $| = 4$
! $\in WORD$, because ! $\in PS$ ($|!| = 1$, '!' regarded as an element of $PS^{\star}$)

But
WHEN$\sqcup \notin WORD$
STOP! $\notin WORD$
!!! $\notin WORD$.

3. THE INPUT FOR THE MORPHOLOGICAL ANALYSIS

Our analysis will accept every $x \in CHAR^{\star}$.

3.1. *Remark.*

$x \in CHAR^{\star}$, $|x| = 0$ $(x = e)$ is a trivial case because there is nothing to analyse.

Let x be a text, $x \in CHAR^{\star}$. If we want to analyse x, we say: " x is the input for the analysis " or for short: " x is the input ".

3.2. *Examples.*

a) $x = \sqcup\sqcup$ *WE'LL* $\sqcup$ *GO* $\sqcup\sqcup\sqcup$ *ON*! $\sqcup\sqcup$
b) $x = 17.23 \sqcup + \sqcup 11.00 \sqcup = \sqcup 28.32 \sqcup - \sqcup 0.9$
c) $x =$ %%%%%%%%%%%%%%%% $\sqcup$ //////////// $\sqcup\sqcup\sqcup$ *ABCDEF* $\sqcup\sqcup$???.

4. THE SEGMENTATION OF THE INPUT

We want to divide some text in a well-defined sequence of words and then take off the blanks.

4.1. *Definition.*

Let *segm*1 be a mapping between $CHAR^\star$ and $(P(WORD))^\star$

$$segm1: CHAR^\star \longrightarrow (P(WORD))^\star$$
$$x \longrightarrow y$$

such that

1) $e \longrightarrow e_p$ (e_p: empty element in $(P(WORD))^\star$)
2) $|x| > 0 \wedge x = a_1a_2a_3...a_k \wedge a_i \in CHAR \ (i = 1, 2, ..., k)$
$\Rightarrow y = \{w_1\}\{w_2\}\{w_3\}...\{w_m\} \wedge w_i \in WORD \wedge$
$w_1w_2w_3...w_m = x \wedge |w_i|$ *maximum* $(i = 1, 2, ..., m)$.

4.2. *Example.*

Let x be: $x := GOOD \sqcup\sqcup\sqcup DAY!$
Then

a) $y' := \{GOOD\}\{\sqcup\sqcup\}\{\sqcup\}\{DAY\}\{!\} \neq segm1(x)$ because $|w_2|$ is not maximum.

b) $y'' := \{GOOD\}\{\sqcup\sqcup\}\{DAY\}\{!\} \neq segm1(x)$ because $GOOD \sqcup\sqcup DAY! \neq x$

c) But $y = \{GOOD\}\{\sqcup\sqcup\sqcup\}\{DAY\}\{!\}$ will fit.
$y = segm1(x)$

4.3. *Remark.*

a) *segm*1 is *ONE-TO-ONE*

Proof.

$$y \in segm1\ (CHAR^\star) \wedge y' \in segm1\ (CHAR^\star) \wedge y = y'$$
$$\Rightarrow \begin{cases} y = \{w_1\}\{w_2\} \ldots \{w_m\} \Rightarrow x = w_1w_2...w_m \\ y' = \{w'_1\}\{w'_2\} \ldots \{w'_{m'}\} \Rightarrow x' = w'_1w'_2...w'_{m'} \end{cases}$$
$$y = y' \Rightarrow w_1w_2...w_m = w'_1w'_2...w'_{m'} \Rightarrow x = x'$$

b) *segm*1 is *ONTO*

Proof.

$\gamma \in (P(WORD))^{\star} \Rightarrow$
1) $\gamma = e_p \Rightarrow \exists\, e \in CHAR^{\star}$: $segm1(e) = e_p$
2) $\gamma \neq e_p \Rightarrow \gamma = \{w_1\}\ \{w_2\}\ \ldots\ \{w_m\}$
Let x be: $x := w_1 w_2 \ldots w_m \in CHAR^{\star}$. Then $segm1(x) = \gamma$

c) Hence *segm1* is a bijection.

4.4. *Definition.*

Let *segm2* be a mapping between $(P(WORD))^{\star}$ and $(P(WORD))^{\star}$

$$segm2\colon (P(WORD))^{\star} \longrightarrow (P(WORD))^{\star}$$
$$\gamma \longrightarrow z$$

such that

1) $e_p \longrightarrow e_p$
2) $\gamma \neq e_p \wedge \gamma = \{w_1\}\ \{w_2\}\ \ldots\ \{w_m\} \wedge w_i \in WORD\ (i = 1, 2, \ldots, m)$
$\Rightarrow \quad z = \{w_{k_1}\}\ \{w_{k_2}\}\ \ldots\ \{w_{k_n}\} \wedge w_{k_i} \notin WORD1 \wedge$
$k_i \in \{1, 2, \ldots, m\}\ (i = 1, 2, \ldots, n) \wedge 1 \leq k_1 < k_2 < \ldots < k_n \leq m$

We change our notation

$$u_i := w_{k_i} \quad (i = 1, 2, \ldots, n)$$

and get

$z = \{u_1\}\ \{u_2\}\ \ldots\ \{u_n\}$.

4.5. *Example.*

$\gamma := \{GOOD\}\ \{\llcorner\!\lrcorner\ \llcorner\!\lrcorner\ \llcorner\!\lrcorner\}\ \{DAY\}\ \{!\}$
$\downarrow$
$z = \{GOOD\}\ \{DAY\}\ \{!\}$

4.6. *Remark.*

a) *segm2* is not *ONE-TO-ONE* (it is *MANY-TO-ONE*)

Proof.

$segm2(\{w\}\ \{\llcorner\!\lrcorner\ \llcorner\!\lrcorner\}) = segm2(\{w\}\ \{\llcorner\!\lrcorner\})$

b) *segm2* is not *ONTO*

Proof.

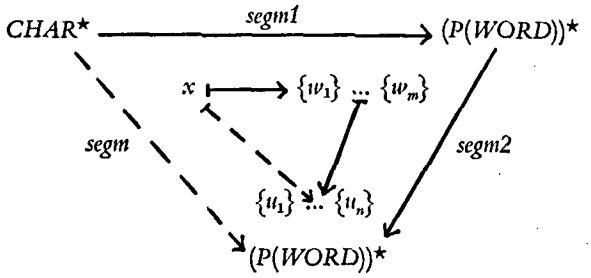
Because of $w_{k_i} \notin WORD1$ there exists no $\gamma \in (P(WORD))^{\star}$ such that $segm2(\gamma) = \{\llcorner\!\lrcorner\}$

4.7. *Definition.*

A *segmentation* of a text is a map *segm* between *CHAR** and (*P* (*WORD*))* such that

$$segm := segm2 \circ segm1$$

The following diagram is commutative:



4.8. *Remark.*

segm is neither *ONE-TO-ONE* nor *ONTO*

Proof.

segm2 is neither *ONE-TO-ONE* nor *ONTO.*

We call every $z \in segm(CHAR^*)$ "a segmented text" or "a segmented input".

4.9. *Example for a segmentation.*

$$x = \llcorner WHAT \llcorner ARE \llcorner\llcorner YOU \llcorner THINKING \llcorner OF \llcorner ?$$

$$segm \left\{ \begin{array}{l} \overline{}\ \overline{} \\ \quad | \quad segm1 \\ \quad \downarrow \\ y = \{\llcorner\} \{WHAT\} \{\llcorner\} \{ARE\} \{\llcorner\llcorner\} \{YOU\} \{\llcorner\} \\ \quad\quad \{THINKING\} \{\llcorner\} \{OF\} \{\llcorner\} \{?\} \\ \quad \overline{}\ \overline{} \\ \quad | \quad segm2 \\ \quad \downarrow \\ z = \{WHAT\} \{ARE\} \{YOU\} \{THINKING\} \{OF\} \{?\} \end{array} \right.$$

5. REMARKS TO THE CONCEPT "LEXICON"

In using the expression "lexicon" all actions identifying and describing words, sentences and texts can be concentrated in a single concept.

In a formal sense any lexicon is a set of "items".

Definition.

$LEX := (\{w, B)_\lambda\}, \lambda \in \Lambda$ (Λ: any index-set)

The pair $(w, B)_\lambda$ is called an "item" of the lexicon.

For every item holds:

a) $w = \{w_1\}\ \{w_2\} \ldots \{w_m\}$, $m \geqslant 1$, $w_i \in WORD2 \cup PS$ $(i = 1, 2, \ldots, m)$

b) w_i $(i = 1, 2, \ldots, m)$ is fixed as a word or punctuation-sign of a language.

c) B is any description of w.

Let IB be the set of all intended descriptions of all sequences of words and punctuation-signs of some language. Then

$$LEX \subset (P(WORD))^\star \times IB$$

such that

$$(w, B) \in LEX :\Leftrightarrow \text{a) b) c)}$$

LEX is a *relation* between $(P(WORD))^\star$ and IB. In general a sequence $\{w_1\}\ \{w_2\} \ldots \{w_m\}$ has more than one description (by homography for example). Hence LEX can't be a map between $(P(WORD))^\star$ and IB.

There are two ways to define a lexicon.

a) *The extensional definition.* All the elements of the lexicon are listed off. In this case we often call such a lexicon a "list".

Examples.

1) $A := \{1, 2, 3\}$
   x is element of $A :\Leftrightarrow x = 1 \vee x = 2 \vee x = 3$

2) w is a noun: $\Leftrightarrow w$ is element of a list of nouns.

b) *The intensional definition.* All the elements of the lexicon are fixed by some common properties.

Examples.

1) $A := \{x \mid x \in IN \wedge 0 < x < 4\}$
x is element of $A: \Leftrightarrow x \in IN \wedge 0 < x < 4$

2) w is a noun: $\Leftrightarrow$ w has some characteristics like prefix, suffix and so on.

3) w is a verb: $\Leftrightarrow$ a part of w (the stem) is element of a list and suffix and prefix have some characteristics.

In both cases (extensional or intensional) the lexicon has the abstract form:

w	description of w
.	.
.	.
.	.

OT (w,B) . . .

6. THE MORPHOLOGICAL ANALYSING STEP

Let $x \in segm(CHAR^*)$

Case 1.

$x = e_p$. There is nothing to analyse.

Case 2.

$\exists\, n \in IN$ such that $x = \{u_1\}\ \{u_2\}\ \ldots\ \{u_n\} \wedge$
$u_i \in WORD2 \cup PS$ $(i = 1, 2, \ldots, n)$.
Let $\{t_1, t_2, \ldots, t_p\} \subseteq \{1, 2, \ldots, n\}$; $p \geqslant 1$
Let k be a map $k: \{t_1, t_2, \ldots, t_p\} \longrightarrow \{t_1, t_2, \ldots, t_p\}$
such that k is *ONE-TO-ONE* and *ONTO*.
Then $(k(t_1), k(t_2), \ldots, k(t_p))$ is a permutation of $(t_1, t_2, \ldots, t_p)$.
We call
$\{u_k(t_1)\}, \{u_k(t_2)\}, \ldots, \{u_k(t_p)\}$ a "subsequence" of $\{u_1\} \ldots \{u_n\}$

Example.

Let $x := \{u_1\}\ \{u_2\}\ \{u_3\}$

Then

a) $\{u_2\}$; $(p = 1)$
b) $\{u_1\}\ \{u_3\}$; $(p = 2)$

c) $\{u_3\}\ \{u_2\}\ \{u_1\};\ (p = 3)$
d) $\{u_1\}\ \{u_2\}\ \{u_3\};\ (p = 3)$

are subsequences of $\{u_1\}\ \{u_2\}\ \{u_3\}$

Definition.

Let T be the set of all subsequences (derived in the above shown manner) of a given $x = \{u_1\}\ \{u_2\}\ \ldots\ \{u_n\}$

Let $t \in T$ be such a subsequence.

Then

a "*morphological analysing step*" related to the subsequence t (for short: mas_t) is a relation between $\{t\}$ and LEX.

$$mas_t \subset \{t\} \times LEX$$

such that

$$(t,\ (w,\ B)) \in mas_t :\Leftrightarrow t = w$$

Case 1. $mas_t \neq 0$

Then we say: we have identified the subsequence t in our lexicon and all related B's are descriptions of t.

Case 2. $mas_t = 0$

Then we say: our lexicon does not (yet) contain the subsequence t. We are not able to give any description of t.

7. THE MORPHOLOGICAL ANALYSIS

The concept of the "morphological analysing step" is related to one and only one subsequence $t \in T$.

The concept of the "morphological analysis" however is more general.

Let $x \in segm(CHAR^*)$, $x \neq e_p$, $x = \{u_1\}\ \{u_2\}\ \ldots\ \{u_n\}$

Let T' be a subset of T $(T' \subseteq T)$ such that to every $\{u_i\}$ $(i = 1, 2, \ldots, n)$ there exists at least one $t \in T'$ which contains $\{u_i\}$.

7.1. *Definition.*

A "*morphological analysis related to* T'" (for short: $ma_{T'}$) of $\{u_1\}$ $\{u_2\}$ ... $\{u_n\}$ is the set of all mas_t such that $t \in T'$.

$ma_{T'} := \{mas_t \mid t \in T'\}$

Remark.

Let $i \in \{1, 2, ..., n\}$

Let t_{u_i} be a subsequence of $\{u_1\}$ $\{u_2\}$ ... $\{u_n\}$ containing $\{u_i\}$.

Let T_{u_i} be the set of all t_{u_i}.

We say that our analysis failed if there exists one $\{u_i\}$ such that $mas_{t_{u_i}} = 0$ for all $t_{u_i} \in T_{u_i}$

In the other case we say that our analysis had been successful.

In general there are more than one T' such that $ma_{T'}$ is successful. It has to be left to the user to fix the sets T' for his special intentions and for his special possibilities.

7.2. *Definition.*

Let AT be the set of all T' (T' defined as above).

A "morphological analysis" (for short: *ma*) is the set of all *mas* such that there exists a $T' \in AT$ with $t \in T'$

$ma := \{mas_t \mid \exists\, T' \in AT \wedge t \in T'\}$

Remark.

Let $z \in CHAR^\star$ be a text such that there exists a $x = segm(z)$ with $x \neq e_p$. Then in practice we say: ma is a morphological analysis of the text z.

8. EXAMPLE FOR A PRACTICAL MORPHOLOGICAL ANALYSIS

This example shows the practice of a morphological analysis of a german text and has indeed been programmed in Bonn to be the basis of the above mentioned syntax-analysis.

Let $x \in segm(CHAR^\star)$, $x \neq e_p$, $x = \{u_1\}$ $\{u_2\}$ ... $\{u_n\}$

Let T' be the following set:

$T' := \{t_i \mid t_i := \{u_i\},\ i = 1, 2, ..., n\} \subset T$

$T' = \{t_1, t_2, ..., t_n\}$

Then holds:

$mas_{t_i} \subset \{\{u_i\}\} \times LEX \quad (i = 1, 2, \ldots, n)$

such that

$(\{u_i\}, (w, B)) \in mas_{t_i} \Leftrightarrow w = \{u_i\}$

This simple case of a morphological analysis we call a '*word-by-word-analysis*' of a given text.

We get

$ma_{T'} = \{mas_{t_1}, mas_{t_2}, \ldots, mas_{t_n}\}$

Each mas_{t_i} $(i = 1, 2, \ldots, n)$ is a set too.

Hence we have to write:

$$ma_{T'} = \{ \begin{array}{l} \{(\{u_1\}, (w_1, B_{11})), \ldots, (\{u_1,\} (w_1, B_{1L_1}))\}, \\ \{(\{u_2\}, (w_2, B_{21})), \ldots, (\{u_2\}, (w_2, B_{2L_2}))\}, \\ \ldots \\ \{(\{u_n\}, (w_n, B_{n1})), \ldots, (\{u_n\}, (w_n, B_{nL_n}))\} \} \end{array}$$

For short we can write:

$ma_{T'}$: $[\{u_i\} \leftrightarrow ((w_i, B_{i1}), \ldots, (w_i, B_{iL_i})) \quad (i = 1, 2, \ldots, n)]$

Or:

$ma_{T'}$: $[\{u_i\} \leftrightarrow (B_{i1}, B_{i2}, \ldots, B_{iL_i}) \quad (i = 1, 2, \ldots, n)]$

Now one can see that the result of a word-by-word-analysis can easily be represented with the following matrix-concept:

$$\begin{pmatrix} u_1, & B_{11}, & B_{12}, & \ldots\ldots\ldots, & B_{1L_1} \\ u_2, & B_{21}, & B_{22}, & \ldots\ldots\ldots, & B_{2L_2} \\ \ldots & & & & \\ u_n, & B_{n1}, & B_{n2}, & \ldots\ldots\ldots, & B_{nL_n} \end{pmatrix}$$

In our syntax-analysis in Bonn a great deal of the morphological analysis is done by word-by-word-analysis. We are successful in describing articles, nouns, adverbs, adjectives and so on, but we have some trouble with our verbs.

In the german language the prefix of some verbs may be found far away from the stem of the verb.

*

Example.

The verbs *zulaufen* and *laufen* are two quite different verbs. We will regard the following three german sentences:

1) *Ein Hund ist mir zugelaufen.*
2) *Lauf mir nur nicht zu.*
3) *Zu ist er mir gelaufen.*

A word-by-word-analysis will succeed only with sentence 1). In 2) and 3) we'll find the verb *laufen* instead of *zulaufen*. That means that we get a wrong description of our verb and a wrong description of *zu* which exists in the german language also without any relation to a verb. Therefore to analyse our verbs a word-by-word-analysis is impossible.

In our analysis we differ between two parts of the lexicon.

The first one allows word-by-word-analysis and is intensionally defined for proper-names, nouns and adjectives and is extensionally defined for all other words without verbs. The extensionally defined part of this lexicon consists at this time of nearly 2000 items.

The second one is our verb-lexicon which is intensionally defined. There exists an extensionally defined verb-stem-lexicon which contains at this moment the stems with their prefixes of nearly 400 german verbs. This stem-lexicon is quickly increasing and is coded in the following manner:

[{*stem*} [description of "stem "]]
[{*prefix*} {*stem*} [description of " prefix stem "]]
[{*lauf*} [description of " lauf "]]
[{*zu*} {*lauf*} [description of " zulauf "]]

We start the morphological analysis with a word-by-word-analysis. If our analysis was successful we have got to each word of some text at least one description. Some of these descriptions may be wrong That's because of the homography and because of the verbs. We can't solve the homography-problem in this early part of the analysis.

If we have identified a word to be a verb we are looking if in the same sentence there exists a word which can be prefix of this verb. If we find a possible prefix the verb gets the descriptions resulting of the prefix as well as the descriptions without this prefix. Working in this way we get a lot of information for the words of our text. Some information is wrong but we can be sure that the right information

is among the descriptions. It is left to the syntax (or maybe to the semantic) to isolate the right descriptions.

Formally we can describe the verb-analysis as a set of mas_t such that $t: = \{u'\}\ \{u''\}$ where $\{u''\}$ has been recognized as a verb and $\{u'\}$ can be every word (without $\{u''\}$) of the same sentence in which $\{u''\}$ exists.

Given some text $\{u\}_1\ \{u_2\}\ \ldots\ \{u_n\}$.

Given a word-by-word-analysis which shows that $\{u_i\}$ may be a verb.

$$T' : = \{t \mid t = \{u'\}\ \{u_i\} \wedge u' \in \{u_1, \ldots, u_{i-1}, u_{i-1}, \ldots, u_n\}\ \}$$

Then holds:

$$mas_t \subset \{\ \{u'\}\ \{u_i\}\ \} \times LEX$$

such that

$$(\{u'\}\ \{u_i\},\ (w,\ B)) \in mas_t \Leftrightarrow w = \{u'\}\ \{u_i\}$$

One might call this procedure a " two-word-analysis ". We can imagine a " three-word-analysis " and so on too, but up to now in our practice in Bonn the morphological analysis consists only of a " word-by-word-analysis " and a " two-word-analysis " in the above shown manner.