## Formal Horphology

Jan HAJIC<br>Research Institute of Mathematical Machines<br>Loretánské nám. 3

11日 55 Praha 1, Czechoslovakia

## Abstract

A formalism for the description of a system of formal morphology for flexive and agglutinative languages (such as Czech) is presented, borrowing some notions and the style from the theory of formal languages. Some examples (for Czech adjectives) are presented at the end of the paper. In these examples, the formalism's rules are used for the phonology-based changes as well, but nothing prevents the use of a separate phonology level (e.g. of the Koskenniemi's two-level model) as a front- (and back-) and for the analysis (and synthesis).

## 1. The Hotivation

Using a computer, the morphological lovel is a basis for building the syntacticom semantic part of any NL analysis. The CL world pays more attention to morphology only after the work /Koskenniemi 1983/ was published. However, as Kay remarked (e.g. in /Kay 1987/), phonology was agtually what was done in /Koskenniemi 1983/. Moreover, the strategy used there is best suited for agglutinative languages with almost one-toone mapping betwean morpheme and grammatical meaning, but slavonic languages are different in this respect.

One of the practioal reasons for formalizing morphology is that although there are some computer implementations using a Czech morphology subsystem (/Hajič,Oliva 1986/, /Kirschner 1983/, /Kirschner 1987/), based on the same sources (/EBSAT VI 1981/, /EBSAT VII 1982/), no unifying formalism for a complete description of formal morphology exists.

## 2. The Pormalina

The terms alphabet, string, concatenation, ..., symbol $N$ (positive integens), indexes * and ${ }^{+}$are used here in the same way as in the formal grammar theory; the symbol $\exp (A)$ denotes the set of all subsets of $A$, $e$ denotes an empty string. Uppercase letters are used mainly for denoting sets and newly defined structures, lowercase letters are used for mappings, for elements of an alphabet and for strings.

Definition 1. A finiteset $K$ of symbols is called a set of grammatical meanings (or simply meanings for short); values from $K$ represent values of morphological categories (a.g. sg may represent singular number, p3 may represent dative ("3rd case") for nouns, etc.).

Definition 2. A finiteset $D=C(w, i) E A^{*} x$ (N y \{OJ) J, where A is an alphabet, is galled a dictionary. A pair (w,i) E D is Galled a dictionary entry, $w$ is a lexical unit and i is called pattern number. In the linguistic interpretation, a lexical unit represents the
notion "systemic word", but it need not be represented by a traditional dictionary form.

Definicion 3. Let $A$ be a firite alphabet, $K$ a finite set of meanings, $V$ a finite alphabet of variables such that $A \cap V=\{ \}$. The quintuple ( $A, V, K, t, R$ ), where $t$ is a mapping $t: V \exp \left(A^{*}\right)$ assignimg types to variables, $R$ is a finite set of rules ( $1, H, u, V, C$ ), where $I$ a $N$ is is afinite set (of labols), C ( C (N (O) C ) is a finite set (of continuations), H $\quad$ ( $\quad$ is a get of meanings belonging to a particular rule from $R, \quad u, v \in(A \cup V)$. is called a controlled rewriting system (CRS); all variables from the left-hand side ( $u$ ) must be present on the right-hand side (v) and vice varsa (rule symmetry according to variables).

Definition 4. Let $T=(A, V, K, t, R)$ ba a CRS. A (simple) substitution on $T$ will be any mapping $q: V \rightarrow A^{*} ; q(V) E t(v)$.

Definition 5. Let $T=(A, V, K, t, R)$ be a CRS and $q$ a ${ }^{\text {simple }}$ substitution on $T$. Mapping $d$ : (A $\left.A^{\prime} V\right)^{*}->A^{*}$ such that $d(\varepsilon)=0 ; d(a)=a$ for a EA; $d(v)=q(v) \operatorname{for} v E V$; $d(b u)=$ $d(b) d(u)$ for $b \in(A v v), u \in(A \vee V)^{*}$ will be called (generalized) substitution derived from $q$.
Comment. The (generalized) substitution substitutes (in a given string) all variables by some string. The same string is substituted for all ocourences of this variable (follows from the definition).

Dafinition 6. Let $T=(A, V, K, t, R)$ be a CRS
 $V)^{*}, i \in N, i{ }^{\prime} \in(N \quad v(0))$. We say that $w$ can be directily rewritten in the state (i,G) to $z$ with a continuation ( $\mathrm{i}^{\prime}, \mathrm{G}$ ') according to meanings $F$ (written as $w(i, G) \quad \Rightarrow[T, F]$ $z(i \cdot G i))$ if there oxist such rule ( $I, H, U, V, C) \in R$ and such Eimple substitution $q$ on $T$, that $i \in I, i^{\prime} E C, H \cap P, G=G \cdot v$ $H, d(u)=w$ and $d(v)=z$, where $d i s$ the substitution derived from $q$.
Relation $\Rightarrow{ }^{*}[P, P]$ is defined as the reflexive and transitive olosure of $=>[T, F I$.
Comment. The CRS is controlled through continuations and labels. After a direct rewriting operation, the only rules that could be applied next must have in their label at least one number from the rewriting operation continuation. Please notice that:

- this operation always rewrites whole words;
- the restriction on the left-hand and righto hand side of a rule that it should be only string (of letters andior variables) is not so strong as it may seem, because no restrictions are imposed on the substitution q. However, to be able to implement the rules in a particular implementation as finite state machines, we shall require $q$ to be defined using regular expressions only.

Dafinition 7. Lat $\mathrm{I}=(\mathrm{A}, \mathrm{V}, \mathrm{K}, \mathrm{t}, \mathrm{R})$ be a CRS and let $n$ be the maximal number from all

Labels Trom all rules from $R$; $n$-tuple $P=$ (p1, ..., pn) will be called a list of patterns; on $T$ (the elements of $P$ ace called patterns; if for every i a mapping pi: exp(K) $x A^{*} \rightarrow \operatorname{sxp}\left(A^{*}\right)$ is defined as $z \in \operatorname{pi}(F, w) \ll$ $\left.w(i, F)=)^{*}[I, F] z(O, C)\right)$.
Comment. The "strange" sets $G$ and $G$ ' from the definition $\sigma$ acquire real meaning only in comection with the definition of patterns; they have a controlling task during pi construction, namely, they check whether all meanings from $F$ are used during the derivation. "To use a meaning k" means here that thare is some rule (I, H, u, v, C) applied in the oourse of derivation from w(i, F) to $z(0,())$ such that $k$ E $H$. Such meaning can then bet removed from $G$ when constructing $G^{\prime}$ (see Def. 7); meanings not from $H$ cannot. Thus, $t 0$ get the empty set in $z(0,[))$ when starting from $w(i, F)$, all meanings from $F$ must be "used" in this sense.
A pattery describes how to construct to a given word w all possible forms according to maanings $F$. In this sense, the notion of pattern does not differ substantially from the traditional notion of pattern in formal morphology, although traditionally, not the constructive description, but just some represencative of such a description is called a patterm.

Definition 0 . Let $D$ be a dictionary over an alphabet $A, T=(A, V, K, t, R)$ a $C R S$ and $P$ a list of patterns on $T$. A quadruple $M=$ ( $A, D, K, E$ ) is called a morphology description on $T$ (M[C]-description).

Definition 9. Let $T=(A, V, K, t, R)$ be a CRS and $N=(A, D, K, P)$ an M[r]-deseription. Set $L$ $=\subset z \in A^{*}$; there ex. we $A^{*}, i \in N, H \wedge K ; z$ pi(H,w)\} will be called a language generated by M[IJ-description M. The alements; of $L$ will be called word forms.
Comment. The term morphology deseription introduced above is a counterpaxt to a description of a system of formal morphology, as used in traditional literature on morphology.
Definition 9 is introduced here just for the purpose of formalization of the notion of word fory, i.e. any form derived from any word from the dictionary using all possible meanings according to M[T].
Definition 10. Let $T=(A, V, K, t, R)$ be a CRS and $M=(A, D, K, P)$ be M[T]-description. The term syrithesis on $M$ is used for a mapping s: $\exp (K) x A^{*} \rightarrow \exp \left(A^{*}\right) ; G(H, w)=\{z ; \theta x . \quad$ i $\epsilon$ N, $i \quad\langle:=n ; z \in p i(H, w) \&(w, i) E D\}$. The texm analysis is used then for a mapping a: $A^{*} \rightarrow \exp \left(\exp (K) \times A^{*}\right) ; a(z)=((H, w) ; z E$ $s(H, w))$.
Comment. According to definition 10 , synthesis means to use patterns for words from the dictionary only. The definition of analysis is based on the syhthesis definition, so it clearly and surely follows the intuition what an analysis is. In this sense, these definitions don't differ substantially from the traditional view on formal murphology, as opposed to Koskenniemi; however, the so-called complex word forms ("have been called") are not covered, and their andalysis is shifted to syntax.

The definition of analysis is quite clear, but there is no procedure contained, capable of actually carrying out this process. However, thanks to rule symmetry it is possible to reverse the rewriting process:

Definition it. Let $T=(A, V, K, t, R)$ be a CRS. Further, let $G$ G' $G^{\prime} K, i \in N, i(E)$ (0)), $z, W \in A^{*}$. We say that under the condition (i', G') it is possible to directly analyse a string $z$ to with a continuation (i,G) (we writez(i', G') $=\langle[T] \quad w(i, G)), \quad i f$ there exists a rule' (I,H,u,V,C) ER and a simple substitution $q$ on $T$ such that $i E I$, i. EC, $G=G^{\prime} y H, d(u)=w a d(v)=z$, where $d$ is the generalized substitution derived from $q$. A relation "it is possible to analyze" ( $=$ '* $^{*}[\mathrm{TJ}$ ) is defined as a reflexive and transitive closure of $=\langle[T]$.

Definition 12. Let $T=(A, V, K, t, R)$ be a CRS and $z \in A^{*}$. Every string $w \in A^{*}, i \in N$ and $F$ n $K$ such that $z(O, C)=\langle[T] W(i, F) i s$ called a predecessor of $z$ with a continuation (i,F). Lemma. Let $T=(A, V, K, t, R)$ be a CRS and we $A^{*}$ a predecessor of string $z \in A^{*}$ with a continuation (i,F). Then $z \in \operatorname{pi}(F, w)$, where pi is a pattern by $T$ (see Def. 7). Proof (idea). The only "asymmetry" in the definition of $\Rightarrow$ as opposed to $=<$, i.e. the condition $H \cap P$, can be solved putting (see
 analysis steps). Then surely Hi n $F$ for every i.
Theorem. Let $T=(A, V, K, t, R)$ be a CRS, $M=$ ( $A, D, K, P$ ) an M[TJ-description, a an analysis by $M$ and $w \in A^{*}$ a predecessor of $z \in A^{*}$ with a continuation (i,F). Moreover, let (w,i) $E$ D. Then $(P, w) E a(z)$.

Proof follows from the preceding lamma and from the definition of analysis.
Comment. This theorem helpa us to manage an analysis of a word form: we begin with the form being analysed $(z)$ and a "continuation" ( $0,(5)$, using then "reversed" rules for back rewriting. In any state w(i,F) during this process, a correct analysis is obtained whenever (w,i) is found in the dictionary. At the same time we have in $F$ the appropriate meanings. Passing along all possible paths of back rewriting, we obtain the whole set a(z).

## 3. An Example

To illustrate the most important features of the formalism described above, we have chosen a simplified example of Czech adjectives (regular declination according to two traditional "patterns" - mladý (young) and jarní (spring), with negation, full comparative and superlative, sg and pl, but only masc. anim. nominative and genitives.

The dictionary:
$\mathrm{D}=\left\{\left(\right.\right.$ nový $\left._{\mathrm{y}} \mathrm{i}\right)$,
(podiy, $\mathbf{y}^{\prime}$ ) $\}$ vile (it has no neg. forms)
The CRS:
CRS $T=(A, V, K, t, R):$
$A=\{a, a ́, b, c, \check{x}, \ldots, \ldots, z, \check{z}, \#\}$
(\# means word separator)
$K=(s g, p l$, comp, sup, neg, masc , nom, acc $\}$
$V=\{-, L, M\}$
$t(-)=A^{*} ; t(L)=\{1, z\} ; t(M)=\{m, n, v\}$
$R=\{$ (seefig. 1) \}


An example of synthesis：we want to obtain s（〔sup，masc，pl，acc\} ypodí́) $\rightarrow$（padiý， 2 ）E $\mathrm{D}_{\mathrm{g}}$ see fig． 2

An example of analysis：we want to obtain a（neinavejší\＃）；see fig． 3

Comment．Better written rules in CRS would not allow for the 4 th alternative in the first step（＂nejnovějsý＂），because＂Ě＂could not be followed by＂ý＂in any Czech word form；however，constructing the other unsuccessful alternatives could not be a priori cancelled－only the dictionary can decide．，whether e．g．＂jnovy＂is or is not a Czech adjective．
Comment on comment．No Cnange in the rules would be necessary if a separate phonology and／or orthography level is used；then，the ＂šy＂possibility，being orthographically im－ possible，is excluded there，of course．

## 4．Conclusion

This formalism will be probably sufficient for Czech（no counter－example to this thesis has been discovered so far）．For inflected words one or two＂levels＂（i．e．， successive rule applications）will suffice，
agglutinative elements（e．g．，adjective comparison）will probably need three to five rules．

References
EBSAT VII（19日2）：Morphemic Analyais of Czech Prague 19 ER
EBSAT VI（1981）：Lexical Input Data for Experimants with Czech，Praha 1981
Koskenniemi，$K_{2}$（1983）：Two－l avel morphology， Univ．of Henlainki，Dupt．of Een．Lingu－ istics，Publications No． 11
Hajič，J．，Gliva，K．（19日G）：Projekt česko－ ruskêhn strajovêhn pr̛akladu，（A Project of Czech to Russian MT System），ins Proceadings of SOFBEM＇E6，Liptovský Jaín
Kirmchner， 2 ．（19日3）：MOSAIC（A Method of Automatic Extraction of Significant Terme from Textel，EBGAT $X$
Kirschner，Z．（1987）：Kirschner，Z．E APAC3－2： An English－to－Czach Machine Translation System，EBSAT XIII
Kay，M．（1987）：Non－Concatenative Finitem State Marphol ogy，Inz Proceedinge of the 3rd European ACL meeting，Coptonhagen， Denmark，April 1997

EBSAT＝Explizite Beschreibung der Spramhe und automatische Textbearbeitung，LK Praha

