LEXICON DESIGN USING A PARADIGMATIC APPROACH

Cristian Dumitrescu

Expert Systems Laboratory
Research Institute for Informatics
8-10 Averescu Avenue,
71316 Bucharest
ROMANIA

Abstract: Based on the MORPHO-2 system, a system that has been developed for managing monolingual lexicons and for lexical processing, some models for representing lexicographic structures and methods for processing them are proposed.

Cristian Dumitrescu graduated from the Faculty of Automatic Control and Computer Science,the Polytechnical Institute of Bucharest in 1986. He started work in the Natural Language Processing Laboratory of the Research Institute for Informatics.In 1992 he joined the Expert Systems Laboratory of the institute.His research interests include natural language processing,mono-and bilingual lexicons management, conceptual lexicons management,knowledge representation and machine translation.

1.Introduction

Modern linguistic theories, no matter the grammar chapter (morphology, syntax, semantics, pragmatics) concerned, hold lexicon of a prominent importance, both practically and theoretically.

Researches are being carried out for identifying models and techniques which could render a lexical dimension to the linguistic generalizations. So, the lexicon can no longer be viewed as a simple list of lexical entries.

The system discussed enables monolingual lexicon handling and incorporates morpho-lexical processes (i.e. word-form analysis and synthesis).

As now approached, the morphological processes obey a paradigmatic morphology (Dumitrescu, 1992), word-forms analysis and synthesis only take into account grammatical endings (which include both desinences and suffixes), and the lexicons handled by the MORPHO-2 system are root- or lemma-oriented.

In order to model paradigmatic morphology and design the lexicon entries, the lexical information is to be properly structured and adequate access mechanisms are to be used to capture linguistic generalizations at lexicon level.

The services provided by the system may be classified according to the following goals: morphological model design, lexical stock building and morpho-lexical processing.The corresponding categories of users will be referred to as: linguist, lexicographer and the target natural language processing system.

The linguist can develop morphological models, making a paradigmatic approach, by means of a proper description language.

We have represented morphological feature bundles as attribute value pairs organized in a hierarchical manner (Dumitrescu, 1991).

It is the lexicographer who gets the new lexicon entries into the hierarchy. The relations of regularity, subregularity and irregularity should be determined for the new word-forms entry and their roots should be specified.

PATR conditions, parameterized macros and macroname overloading provide the specifications for a syntactical description of a lexical entry.

As far as the linguist and lexicographer are concerned, to express the lexicon, the system offers a lexical representation language.

By compiling the available lexical information, optimal structures will be generated for morpho-lexical processing.

For the target natural language processing system, which benefits the morpho-lexical processes, MORPHO-2 is a lexical information retrieval system.

2. Morphological Model Design

The linguist will have to proceed on a step-by-step elaboration of the morphological model and define the following:

a) feature value domains

b) categories, subcategories, features and their values, in a hierarchical manner

c) paradigmatic descriptions

d) lemma - entry correspondence, for each paradigmatic description

e) feature specification defaults associated with each paradigmatic description

f) inflectional paradigms and root detection rules.

The feature value domain specifications allow for feature value checking during the model compilation and for the use of (ANY) value as abbreviation of the values of an entire domain . Feature value domains also help in analysing the word- forms. The system tries to compact the morphological descriptions of those word- forms occurring in various slots of a paradigmatic description.Hence ,some features are given more values (the ANY value will be used in case of full domain values). This usually happens with feature specifications of the terminal nodes in the feature hierarchy (Figure 1).

Some examples of feature value domain specification { **NMB:G, PL; PER: 1,2,3.** } are given in the following.A hierarchical description of features is possible by correlating several feature specifications. A feature specification is given under the form of a (feature: value +) pair, where feature and values are atomic. A hierarchical description using several simple (feature:value)pairs is called paradigmatic description. The morphological model of the Romanian language is hierarchically described in Figure 1, where it takes the form of an incomplete tree.

Each non-terminal node contains a single feature specification. The leaf nodes may contain one or more features specifications. By applying the successor's selection criteria to a non-terminal node, a distinction can be made between CHOOSE nodes (when only one successor is selected) or FOREACH nodes (when an individual selection of each successor is required). A curve intersecting the emerging edges of FOREACH node is presented in Figure 1. By traversing the tree across the longest path which starts from the root node, through CHOOSE nodes only, a paradigmatic description selector is obtained (e.g. **CAT = NOUN& SCAT = COMMON & GEN = FEM, CAT = VERB**).

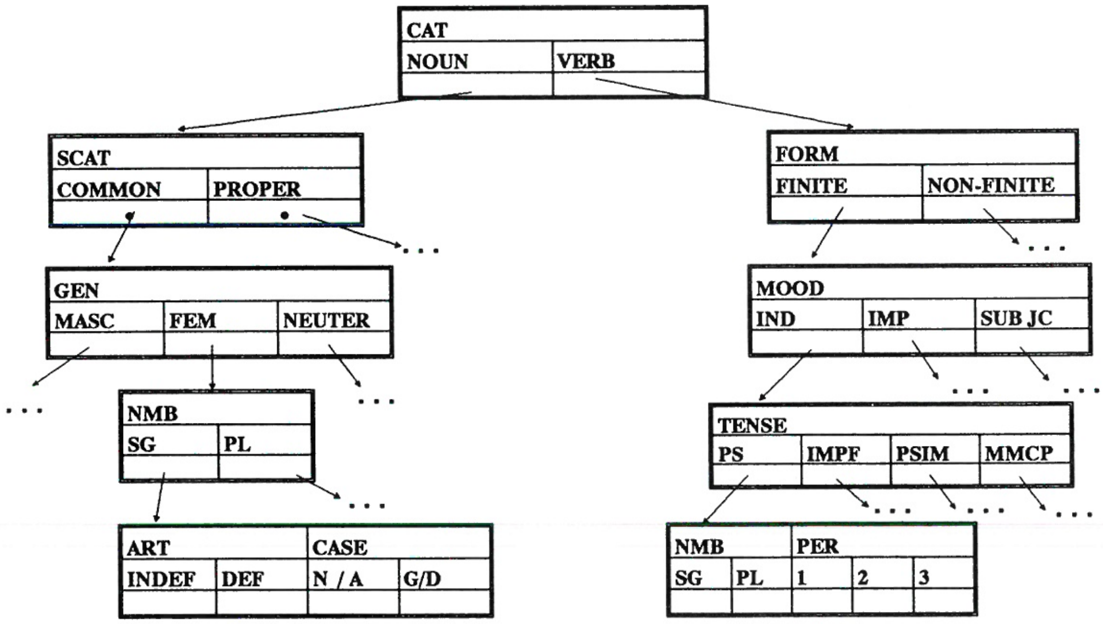

Figure 1. Hierarchical Description of Features

A morpho-lexical acquisition scenario is used in describing a leaf node. A scenario entry (from now on referred as a slot) corresponds to a point of the paradigmatic description space.

The Figure 1 specifications generated paradigmatic descriptions, i.e. feminine common noun declension and verb conjugation, as presented in Figure 2.

A morphological model where the lemma-entry relation is paradigmatically described includes the points specification in the paradigmatic description spaces, circumscribing the lemma field in the lexicon entry. Thus, the lexical transfer takes place at the proper lexical level.

CAT=NOUN&SCAT=COMMON&GEN=FEM

NMB = SG

ART	CASE	WORD_FORM
INDEF	N/A	
INDEF	G/D	
DEF	N/A	
DEF	G/D	

NMB = PL

...

...

CAT = VERB

FORM = FINITE

MOOD = IND

TENSE = PS

NR	PER	WORD_FORM
SG	1	
SG	2	
SG	3	
PL	1	
PL	2	
PL	3	

TENSE = IMPF

...

TENSE = MMCP

MOOD = SUBJC

...

Figure 2. Morpho-lexical Acquisition Menus

(Feature: value) pairs-default inheritances from the corresponding slots are available to the selectors of the descriptions allowing for feature specification defaults. In our example the following association is possible:

(CAT=VERB --- (PER 1, 2, 3).

How the system is to build up inflexion paradigms and root detection rules is eventually described by the morphological model. The linguist will join several paradigmatic ending families to each paradigmatic description, letting the system choose among them when building inflectional paradigms. 136 inflectional paradigms have been identified for the Romanian language.

Based on the inflectional paradigms, the system will determine root detection rules and word-form generation.

A root detection rule will be the following:

```
<inflection>:=(<inflectional-paradigm>
               <slot-number>)
```

explained as:

a) **if** a word ends in <inflection> **then**

- the root is what remains from the word after dropping out the <inflection>
- the root belongs to the <inflectional-paradigm>
- the contextual information corresponding to the current word is given by <slot-number>

b) **if** a root belongs to the <inflectional-paradigm> and is used in the context given by <slot number> **then**

- the word is obtained by concatenating the given root with the <inflection>.

The lexicographer's interface strictly depends on the linguist's interface specifications, a large part of the former automatically deriving from the latter.

3. The Lexicon Entry

The lexicon entry has the following formal structure:

```
<lexicon-entry>::=
(<lemma>
    (<paradigmatic-description-selector>
      <inflectional-paradigm>
```

(< morphological-description > < root >)*

(< syntactic-description > < semantic-description)*)*)

By a the standard form of a word is referred: indefinite nominative singular for nouns, short infinitive for verbs, etc.

The paradigmatic-description-selector field has the already established meaning.

The morphological model compilation brings about an hierarchy of features and the specified inflectional paradigms are associated with the paradigmatic descriptions generated by that hierarchy. When defining a new lexicon entry, the lexicographer refers the hierarchy (by paradigmatic description selectors) and the inflectional paradigms labelling.

A straightforward, but, on the other hand, quite inefficient way of representing word-forms in a paradigmatic description is by merely filling in the corresponding slots. Redundancy can be limited by providing the inflectional paradigms and the corresponding roots.

Furthermore, a non-monotonic inheritance mechanism will help in defining the regularity, subregularity and irregularity relation of the word-forms (Gazdar, 1988), (Evans and Gazdar, 1989).

The(morphologic-description root)* fields combine the current roots in the paradigmatic description referred to by the selector.

In fact, associations are yielded by rules of the form:

[path^1] ⇔ root^1

[path^2] ⇔ root^2

....

[path^n] ⇔ root^n

where each path starts at the top of the subtree defining the paradigmatic description.

Given the above feature hierarchy, let us consider the feminine common noun description (Figure 3).

The morphological model makes use of an association:

(CAT = NOUN & SCAT + COMMON & GEN + FEM) -- (CASE N/A/G/D/V)

which results in default inheritances for the feature CASE.

In this context the root association rules for the lemma FEMEIE (woman) will be as specified below:

(FEMEIE

[CAT = NOUN & SCAT = COMMON & GEN + FEM]

INFLPR26

[NMB = SG] ⇔ FEMEI

[NMB = PL] ⇔ FEME)

...

CAT = NOUN&SCAT = COMMON&GEN = FEM

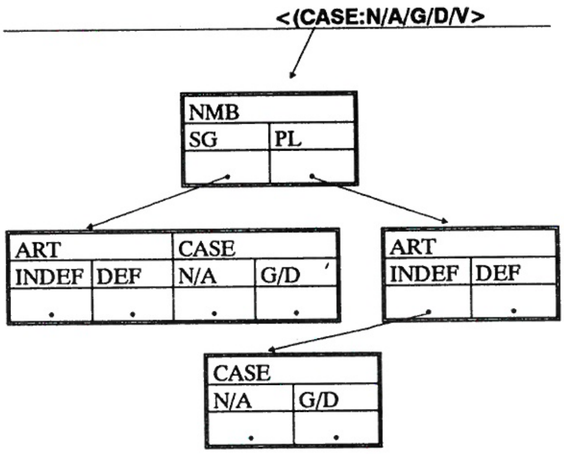

Figure. 3 Hierarchical Description of Feminine Common Noun

More precisely, a rule of the form:

(path^i) ⇔ root^i

has tHe following double meaning:

a) **if** path^i has an associated root^i **then**

* root^i is the default inheritance for the slots reached through path^i

b) **if** root^i is associated with path^i **then**

* root^i inherits the morphological features bundled together by selector, feature specification defaults and path^i.

One can easily get total or partial regularity by applying such rules. The same mechanism applies for exceptions.

Thus, given the rules:

[path^{i}] $\Leftrightarrow$ root^{i}

[path^{j}] $\Leftrightarrow$ root^{j}

so that path^{i} $\subset$ path^{j} (path^{j} is an extension of path^{i}) then root^{j} overwrites root^{i} in the path^{j} slots.

The following examples use this technique:

(FATA

([CAT=NOUN & SCAT=COMMON & GEN=FEM]

INFLPR30

([] $\Leftrightarrow$ FET

[NMB=SG & CASE=N/A] $\Leftrightarrow$ FAT)

...)

(FEMEI

[CAT=NOUN & SCAT=COMMON & GEN=FEM]

INFLPR26

([] $\Leftrightarrow$ FEME

[NMB=SG] FEMEI

...]

There are three types of syntactic information common to all word-forms of a lexical entry (in fact, of the sub-entry uniquely identified by lemma and paradigmatic description selector):

a) part of speech

b) valency models (how to combine with other words, for instance verb transitivity)

c) certain inherent feature specifications which are syntactically relevant (e.g. gender for nouns).

In our approach, the information referred to by a) and c) is provided by the paradigmatic description selector (for instance **[CAT=NOUN & SCAT=COMMON & GEN=MASC], [CAT=NOUN & SCAT=PROPER & GEN=FEM].**

The b) information type is the one contained by the <syntactic - description> field. The restrictions of co-occurrence with other words (or phrases) have been revealed when extending the category-valued features used in PATR-like representation (Shieber, 1986, Estival, 1990) to the Romanian language.

As an example let us consider the transitive verb ARATA (*to show*).

Lemma: **ARATA**

Selector: **CAT= VERB**

PATR *conditions*: <subcat> = [OBJ SUBJ]

<SUBJ cat> = np

<SUBJ head case> = n

<OBJ cat> =np

<OBJ head case> = a.

By means of the mentioned syntactic constraints, the transitive verb ARATA places in the complement subcategory a nominative subject np and an accusative object np.

The lexical representation language should allow that those feature structure aspects which are common to transitive verbs are expressed but once and are not repeated for each individual lexical entry. In (Shieber, 1986) techniques using syntactic macrodefinition are described in order to include parts of the feature structures, shared by several lexicon entry classes. A lexical entry subject to this improvement will, besides its own syntactic constraints, contain syntactic macrodefinition names to be expanded upon request only.

The approach enables a compact lexical entry representation as well as capturing of the available lexical generalizations, given that macrodefinitions can be hierarchically ordered.

One extension of the PATR conditions is the special attribute this (hereafter '*') at the top of their path description. It refers the lexicon entry under analysis. Actually, our intention is that, given a word-form, a complete feature structure is obtained by unifying the descriptions yielded by the lexicon entry, if lexically analysed.

Let's say that the lexicon entry for the lemma AJUNGE (*to get to*) contains the following PATR conditions:

* head agreement per = * per (1)

```
* head agreement nmb = * nmb
```

and if the morphological analysis of the word-form AJUNG leads to:

```
* cat = verb
* mood = ind
* tense = ps                              (2)
* nmb = sg
* per = 1
```

then (1) and (2) can be unified and thus, our feature structure can be enriched.

By a new extension, described below, atomic disjunctive values (e.g. n/g) and list values (e.g [SUBJ]) are specified.

```
Macro InTrans:
    < * subcat > =[ SUBJ ]
    InTran.
Macro InTran:
    < SUBJcat > = np
    < SUBJhead case > = n/g
    < * head agreement per > = < * per >
    < * head agreement nmb > = < * nmb >
    < * head agreement > = < SUBJ head agreement >.
```

Using parameterized macros and macro name overloading, the valency models for the Romanian transitive verbs may be easily expressed as Trans(np), Trans(pp), Trans(np/pp/pp_pron), etc.

```
Macro Trans(np):
    < *subcat > =[ OBJ SUBJ ]
    InTran
    < OBJcat > = np
    < OBJ head case > = a.
Macro Trans(pp):
    < * subcat > =[ OBJ SUBJ ]
    InTran
    < OBJ cat > = pp
    < OBJ head case > = a
    < OBJ head pform > = pe/la.
Macro Trans(pp_pron):
    < * SUBCAT > =[ [ OBJ1 OBJ2] SUBJ ]
    InTran
    < OBJ1 cat > = pp
    < OBJ1 head case > = a
    < OBJ1 head pform > = pe
    < OBJ2 cat > = pron
    < OBJ2 head case > = a.
```

A peculiarity of the Romanian language is that of doubling the direct object revealed by the latter macro. For instance, in the next sentence the direct object (**pe Ion**) is doubled by accusative personal pronoun (**L-**):

L-am vazut **pe Ion**.

*

I **him** have seen **John**.

but the "two" direct objects refer to the same object and therefore only one valency is required.

The lexicographer can provide one semantic description for each syntactic description. In our opinion, the semantic description of an analysis and generation lexicon (as the one presented here) should mediate between a given natural language and the meaning representation language.

The semantic representation field has a PATR-like form too (semantic macro-definitions included).

On defining the PATR macros, the lexicographer must select the restricted path out of the semantic description fields which the meaning representation reference index will be built for.

Example:

```
Restrict:
    head sem arg
...
Patr Macros:
...
Macro SemArg(Arg):
    head sem arg = Arg.
Macro Sem (Arg):
    SemArg(Arg)
    Finite(*)
```

Mood(ind)

head sem tense = * tense

head sem form = real.

The semantic representation references label different case-frame structures placed in a generic-specific hierarchy.

Lexical ambiguity, marked by more than one label to a lexical entry, is due to either category ambiguity (e.g. noun vs. verb) or polysemy and homonymy. To solve the latter type of ambiguity a detailed meaning and contextual analysis should be carried out. Additional mechanisms are therefore needed. Thus, the actual semantic descriptions are stored in a database separately from the rest of the lexicon (Nirenburg, 1987) and managed independently of MORPHO-2.

An example describing a complete lexicon is given below:

(AJUNGE

(CAT=VERB 48

([] **<=>** AJUNG

[FORM=FINITE MOOD=IND &

TENSE=PSIM/MMCP] **<=>** AJUNS

[FORM=NON_FINITE & MOOD=PART] AJUNS)

(([InTrans Sem(s_int)])

([Traus(np/pp/pp_pron)] [Sem(a_agent/s_des)])

([Inf] [SemArg(a_name/s_name)])

([Part] [SemArg(s_pos/s_des)])

([Ger] [SemArg(s_act)]))))

4. Morpho-Lexical Processing

The target natural language processing system is the beneficiary of the morpho-lexical processes executed under MORPHO-2.

For a given sequence of words, a morphological analyser yields valid root-ending pairs. We call a valid (rooti, inflectionj) word segmentation the one obeying the following:

- **DP** be the word paradigmatic description
- **DIj** be the inflectionj morphological description after applying root detection rules.
- **DR1, DR2,...,DRk** be the paths attached to the roots in **DP** (as stated by root association rules in **DP**)

then

- **DIj** is a full path in **DP**
- **DRi** is the longest path in the (**DR1, DR2,...,DRk**) set, such that **DIj** is an extension.

In order to efficiently retrieve the corresponding lexical information, lexical entries are indexed on roots and inflectional paradigms (which are in turn indexed on endings).

For the obtained root-ending pairs, the system will then build the morpho-lexical atoms.Based on the unambiguous atoms, an attempt will be made at constructing minimal DAGs (**Directed Acyclic Graph**) satisfying the atoms included set of PATR conditions.

4.1. Morpho-Lexical Atoms

The structure of a morpho-lexical atom is given below:

```
(<root>
  (<lemma>
      (<paradigmatic-description-selector>
      (<morphological-description>*
      (<syntactic-description>
       <semantic-description>)*)*)*)
```

A morphological description contains both contextual and context- free information. The former is obtained from ending analysis and the latter from the lexicon entry corresponding to the root. The information on the other fields in the atom structure is also supplied by the lexicon entry corresponding to the root.

A morpho-lexical atom is given as an example of the **OMUL** (*the man*) word-form.

```
(OMUL
(OM( [[cat noun][ scat common][ gen masc]]
    ([[nmb sg] [ art def ][ case [ / n a ] [ per 3 ]] )
    (([ Noun]  [ SemArg(man)] )))))
```

The morphological congruence and the root retrieval within the lexicon provide the key for classifying the morpho-lexical atoms as unambiguous,ambiguous and undetermined.

The unambiguous morpho-lexical atoms associate the analysed word with a single lemma.

With the same paradigmatic description selector, the system will attempt a compaction in case that a root which corresponds to one lemma has more possible morphological descriptions.

Compaction technique refers the feature value domains contained in the object schema (resulted from the morphological model compilation).

Example:

```
(LATR
(LATRA
([ CAT = VERB ]
([ FORM = FINITE & MOOD = IND & TENSE
= PS & NMB = SG & PER = 9 ]
 [ FORM = FINITE & MOOD = IND & TENSE
= PS & NMB = PL & PER = 9 ]
 [ FORM = FINITE & MOOD = IND & TENSE
= IMPF&NMB = SG & PER =
...

 [ FORM = FINITE & MOOD = IMP & PER =
2 & NMB = SG ]
)
...
)
```

Compaction:

```
{ [ FORM = FINITE & MOOD = IND &
  TENSE = PS & NMB = SG & PER = 9]
  [ FORM = FINITE & MOOD = IND &
  TENSE = PS & NMB = PL & PER =9] }

[ FORM = FINITE & MOOD = IND &
TENSE = PS & NMB = * & PER = 9]
```

The ambiguous morpho-lexical atoms derive from words to which several lemmae may be attached. A root association with several lemmae is possible because of the ambiguity of category (e.g. noun vs. verb), on the one hand, and because of the apparent homography generated by the absence of prosodic markers in the Romanian language (modele, modele, àcele, acéle, modul modul, etc.),on the other hand. Possible interpretations are ordered in such a way that those coming from shorter roots (i.e. longer ending) - get priority.

The undetermined morpho-lexical atoms correspond to words with no entry in the lexicon. The atoms structure in this situation will be the following:

```
(UNKNOWN <unknown-word>
(<possible-root> <(morphologic-description>*)*)
```

The unknown word is associated with all legal segmentations and each segmentation will be assigned morpho-lexical information deduced from the identified endings.

The system generated analysis of the unknown word **PICIORE** (*legs*) is made. Default feature specification and feature value compaction are illustrated.

```
(UNKNOWN PICIOARE
((PICIOAR
([[ cat verb ][ form finite][ mood ind ][ tense ps][
nmb *][ per3]]
[[ cat verb][ form finite][ mood ind][ tense psim][
nmb sg][ per 3]]
[[ cat verb][ form finite][ mood imp][ per 2][
nmb sg ]]
[[ cat verb][ form finite][mood subjc][ nmb *][
per 3]]
[[ cat verb][ form non-finite][ mood inf][ per *][
nmb *]]
[[ cat noun][ scat common][ gen masc][ nmb sg][
art indef][ case *][ per 3]]
[[ cat noun][ scat common][ gen fem][ nmb sg][
art indef][
  case [ / n a g d]] [ per 3 ]]
```

[[cat noun][scat common][gen fem][nmb pl][art indef][case *][per 3]]

[[cat noun][scat common][gen neuter][nmb *][art indef][case *]

[per 3]]))))

A further compaction could be the elimination of the ANY value feature specification. In this case, the difference between the ANY features and the non- allowed features is no longer possible. Therefore, such specifications are really helpful.

5. Implementation

The MORPHO project, dated back as 1987, has produced a first result, a prototype now available on a PDP-11 compatible computer. A second version of the system presented in this paper is being developed under C+ + on an IBM-PC compatible.
The lexicon entry architecture as well as the relation type between its fields are the same for all lexicons handled by MORPHO-2 and are not accessible to the user definition. The access methods to each entry field, the relations among entry fields as well as among different entries are directly controlled by the system.
Such restrictions are not to be interpreted as system limitations but as required by a disciplined approach to the lexicon building process.
The use of multilists and the handling of variable length records have been claimed by the lexicon entry structure. Indexing lexicon by means of prefixed virtual B + tree and an optimal grouping of data about morpho-lexical processing have determined an average response time of lexical processes,fully independent of the lexicons size (for more details on performance analysis see(Tufis and Dumitrescu,1990)).

REFERENCES

DUMITRESCU,C., MORPHO-2 Design and Development Environment for Monolingual Lexicons, ROMANIAN JOURNAL OF INFORMATION TECHNOLOGY AND AUTOMATIC CONTROL, Vol.1, No.2, Bucharest, 1991, pp.23-27.

DUMITRESCU,C., Paradigmatic Morphology Modelling and Lexicon Management with MORPHO-2,Proceedings of the 5th EURALEX International Congress,Tampere,1992.

DUMITRESCU,C., MORPHO-2 Reference Manual,Research Institute for Informatics, Bucharest, 1992.

ESTIVAL,D., ELU User Manual, ISSCO, Geneva,1990.

EVANS,R. and GAZDAR,G., Inference in DATR,Proceedings of the 4th Conference of ECACL,Manchester,1989,pp.66-71.

GAZDAR,G., The Organization of Computational Lexicons,Cognitive Science Research Paper,The University of Sussex,Brighton,1988.

GAZDAR,G. and MELLISH,C., Natural Language Processing in PDP-11:

An Introduction to Computational Linguistics,ADDISON WESLEY, Wokingham, England,1989.

KILBURY,J. et al, DATR as a Lexical Component for PATR,Proceedings of ECACL '91, Berlin, 1991,pp.137-142.

NIRENBURG,S. and RASKIN,V.,The Subworld Concept Lexicon and the Lexicon Management System,COMPUTATIONAL LINGUISTICS,Vol.13,Nos.3- 4,1987,pp.27o-289.

SHIEBER,S., An Introduction to Unification-based Approaches to Grammar, CSLI/SRI International, Stanford, CA.,1986.

TUFIS,D. and DUMITRESCU, C., MORPHO-A Dictionary Management System,Proceedings of the 13th International Seminar on DBMS,Mamaia,Romania,1990,pp.174-182.