Lakoff, G. & Johnson, M. 1980. *Metaphors we live by*. The University of Chicago Press.

Lenz, R. 1895-97. *Estudios Araucanos*. Anales de la Universidad de Chile XC, XCI, XCII, XCIV, XCVII, XCVIII.

Lenz, R. 1944. *La oración y sus partes*. 4th ed. Santiago de Chile: Editorial Nascimento.

Lyons, John. 1968. *Introduction to theoretical linguistics*. London: Cambridge University Press.

de Moesbach, E.W. 1930. *Vida y costumbres de los indígenas Araucanos*. Santiago: Imprenta Cervantes.

Rivano, E. 1987. 'Funciones sintácticas en Mapudungu'. *Revista de Lingüística Teórica y Aplicada* 25, 19-26. Universidad de Concepción (Chile).

Rivano, E. 1988. 'Morphosyntactic functions in Mapudungu'. *Revista de Lingüística Teórica y Aplicada* 26, 57-90. Universidad de Concepción (Chile).

Rivano, E. 1989. 'Modeling personal relations'. Unpublished.

Salas, A. 1979. *Semantic ramifications of the category of person in the Mapuche verb* (Spanish text). Ph.D. Thesis, State University of New York at Buffalo. University Microfilms International.

Salas, A. 1983. 'Dos cuentos mitológicos Mapuches: El Sumpall y El Trülke Wekufü. Una perspectiva etnográfica'. *Acta Literaria* 8, Concepción.

Salas, A. 1984. *De la etnología a la literatura, de la literatura a la etnología. Cultura, hombre y sociedad*. Revista de ciencias sociales y humanas. Pontificia Universidad Católica de Chile.

Lund University, Dept. of Linguistics
Working Papers 35 (1989), 169-190

Morphology in Referent Grammar and in the Automatic Translation System SWETRA

Bengt Sigurd, Barbara Gawrońska-Werngren and Mats Eeg-Olofsson

Abstract
Most of the work in Referent Grammar so far has focused on syntax, but it is clear that the syntactic rules must rely on some kind of lexical information about the form, meaning and category of the words to be used in the rules. As the lexicon cannot include all possible forms, at least not in such languages as Swedish, Russian and Georgian, some kind of morphological rules are needed. Such rules also reflect the morphological competence of the language users, as demonstrated e.g. when new loan-words are to be inflected. This joint paper is a discussion of the problems met when morphological rules are to be integrated in Referent Grammar and when morphological rules are to be localized in the whole multilingual translation system SWETRA.

INTRODUCTION

Referent grammar (RG; Sigurd 1987) is a type of generalized phrase structure grammar enriched with functional representations. Its noun phrases also include numbered referent variables which has given the grammar its name. Referent grammar is written directly in the Definite Clause Grammar formalism (supported by most Prolog programs) and can therefore be run and tested both in analysis and generation directly on computers. Referent grammar is used in the automatic translation project SWETRA (Swedish Computer Translation Research, supported by The Swedish Research Council for the Humanities and Social Sciences) and extensive grammatical modules for English, Swedish and Russian have been implemented (Sigurd & Gawrońska-Werngren 1988).

The RG grammar rules can analyze a sentence and give the equivalent functional representation(s) or generate a sentence if given a functional representation. The RG analysis also gives information about the mode and the focused constituent of the sentence. The following are some simplified RG rules which may be found in a grammar module analyzing and generating Swedish. The rules can analyze and generate such sentences as: *Idag kom hunden* (literally: 'Today came the dog').

```
sent(d,B1,s(subj(B3),pred(B2),advl(B1))) --> adv(B1),vi(B2),nps(B3).
adv(B1) --> [A],{slex(A,B1,adv,_,_,_,_,_,_,_,_)}.
vi(B2) --> [V],{slex(V,B2,v,vi,_,_,_,_,_,_,_)}.
nps(B3) --> [N],{slex(N,B3,n,_,_,_,_,nom,_,_)}.
```

The first rule states that a Swedish declarative sentence may include an adverb (adv), an intransitive verb (vi) and a subject noun phrase (nps) in that order. The mode is marked by 'd' in the first slot of the basic unit (Prolog predicate) 'sent'. The equivalent standardized functional representation is given in the third slot of 'sent'. This functional representation, s(subj, pred, advl), would also cover cases with different surface word order, e.g. *Hunden kom idag* 'The dog came today', where *hunden* is focused, however. Differences in the focused (topicalized, preposed) constituent are registered in the second slot of 'sent'. In our first example the (meaning of the) adverb (B1) is placed in the focus slot.

The second rule states that a form (A) is an adverb (and thus can participate in the first rule) if it is a string of letters (included in []), and if this string is registered as a form in the Swedish lexicon (slex) with the category feature 'adv'. The third rule shows that an intransitive verb is marked by 'vi' in the lexicon, and the fourth rule shows that subject nouns are marked by 'nom'. The variables B1, B2, B3 denote the word meanings, which the grammatical rules carry into the functional representations instead of the forms (strings) A, V, N as can be seen.

As all syntactic rules, the syntactic RG rules rely on the lexicon, the store where the form and meaning of words are related and their syntactic possibilities are indicated. The RG format of all lexical items is: lex(A,B,C,D,E,F,G,H,I,J), where the variables A, B, C, etc. vary with the word. The Swedish lexicon has a prefix 's' (slex), the English lexicon has the prefix 'e' (elex), and the Russian lexicon has the prefix 'r' (rlex). The graphic form (entry) is always found in the first slot (A) of lex, the meaning (B) in the second slot, and the word class (C) in the third slot. The information in the later slots vary with the word class. For nouns, later slots contain information about definiteness, number, gender, sex, case, declension class and semantic type. The following are some sample Swedish lexical entries, where the meaning appears (with 'm') in 'machinese English' in the second slot:


```
slex(hunden,m(dog,sg),n,def,sg,re,_,nom,d2,_).
slex(barn,m(child,sg),n,indef,sg,ne,_,nom,d5,_).
slex(kom,m(come,past),v,vi,_,_,_,_,c5,_).
slex(idag,m(today,_),adv,_,_,_,_,_,_,time).
slex(snabb,m(quick,_),a,indef,sg,re,_,_,_,manner).
```

SOME PRELIMINARY MORPHOLOGICAL CONSIDERATIONS

If all the word forms of a language could be included in the lexicon, no morphological rules would be needed; studies of form and meaning relations between the words of the lexicon would only be of secondary interest. But in accordance with the linguistic tradition, RG considers certain forms as secondary (derived) in relation to other primary (basic, core) forms. Intuitively, we see some forms as being built from others by adding an affix, e.g. *barns* 'child's' as being derived by the addition of an *s*. In other cases one form can be seen as constructed from another by changes of certain sounds (letters), e.g. *came* from *come*.

There are several good reasons for the traditional approach, which survives in most modern formal grammatical theories, if they care to treat morphology – which is by no means the rule (cf. Anderson 1982, Hammond & Noonan 1988). One reason is that the speakers always know how new or unknown words are to be inflected. Speakers must have an internalized system allowing them to assume e.g. that the genitive of a personal name *Vilibrat* must be *Vilibrats* in Swedish. A Swede also knows that the past tense of the verb pronounced as *sejva* (loan from the English computer term *save*) must be *sejvade*, although he has never heard the word (and the Swedish Academy has not said anything on the question). Furthermore, a Swede knows that if there is an adjective A, its neuter form should have an additional *t* and that this form also can be used as an adverb. There should thus also be rules stating how one form can be used as another category. In terms of the lexical items presented above, a Swede knows that if there is the lexical item

```
slex(snabb, m(quick,_),a,indef,sg,re,_,_,_,manner),
```

there is also a potential (derived, secondary) lexical item

```
slex(snabbt,m(quick,_),a,indef,sg,ne,_,_,_,manner)
```

and a potential (derived, secondary) lexical item

slex(snabbt,m(quickly,_),adv,_,_,_,_,_,_,manner).

The last item could obviously be used in generating sentences by our little grammar, e.g. *Snabbt kom hunden* (literally: 'Quickly came the dog').

With an even more general, bidirectional or non-directional formulation (cf. Eeg-Olofsson 1988), the speakers' morphological knowledge can be seen as only a net of relations between forms. A Swede knows that if a word is a noun its genitive normally ends in *s*, but also that if he finds a noun ending in *s*, its nominative form probably lacks this *s*. If he has heard about *Vilibrats bil* 'Vilibrat's car' he assumes that the nominative form of the name is *Vilibrat*. Due to this knowledge he is always able to establish the base forms of a lexicon, if he wants, but his morphological knowledge can be seen as just a net of morphosemantic relations.

We also know that a lexicon can never be finite; new words are continually added to living languages due to the development of social life, culture and technology. Therefore there must be rules which allow the users of a language to derive the full set of morphologically related forms spontaneously, whenever a new word appears, and in whatever morphological form it appears, when it is first met.

There are certain psycholinguistic reasons for assuming a lexicon with primary forms and rules deriving other forms. Such a lexicon can be much smaller, take less mental space, and be more manageable. These reasons are made concrete in computer implementations of natural language processing (NLP) systems, as the work space of the computer is restricted and the computer simply cannot include an infinite number of lexical items. Storing all word forms with their meanings seems redundant.

Furthermore, there seem to be reasons for assuming that forms are only evoked or generated when needed in the use of the language. In the analysis (parsing) of Swedish sentences, possible adverbial word forms are needed when analyzing the beginning and end of Swedish sentences, and possible genitive forms are needed when one is looking for the prenominal modifiers of a noun. In synthesis (sentence generation), when one knows what to say, the particular inflected word form is not needed until one has decided what to say and decided on the phrase structure of the sentence.

The problems of morphology in RG and in SWETRA can be summarized as follows, assuming that a lexicon with all inflected and derived word forms is not available:

1. Which (primary) forms should be included in the (core) lexicon? (the primary form problem)
2. Which rules and arrangements are needed to derive all the possible forms from the primary ones in the lexicon? Can non-directional rules be designed and used both in analysis and synthesis? (the morphological rule problem)
3. Which (additional) forms should be included in the lexicon as not being derivable by rule? (the irregularity problem)
4. Where in the grammar or in the whole translation procedure should morphological rules be called upon? (the morphological localization problem)

We will first treat ways of integrating morphological rules in the syntactic rules of Referent Grammar irrespective of the use of the grammar in an automatic translation system. We will then discuss where morphological processes of different types can be applied in the whole SWETRA system. Examples from Swedish and Russian (in a special transliteration) will mainly be used.

1. INTEGRATING MORPHOLOGICAL RULES IN REFERENT GRAMMAR.

The inflectional component of RG suggested below is based on the following principles:

1) bidirectionality of the morphological knowledge: both recognition and production of a certain word form is achieved by means of the same rule or set of rules and/or lexically stored information. The difference between form analysis and form generation lies in the way of processing, not in the morphological knowledge itself.

2) differentiated processing: the morphological analysis starts with identification of phonological connections between word forms, while in the generation process, the starting point is the meaning for which the proper word form is sought.

3) local syntactic triggering of inflectional rules: when processing a certain syntactic structure (e.g. a noun phrase), we activate primarily the part of the morphological knowledge which is necessary for the comprehension and production of that particular structure.

4) predictive triggering: identification of the morphosyntactic properties of a certain word form may give some indications about possible morphological rules needed for recognizing another form. When e.g. a word form has been recognized as an adjective having the features fe, sg, acc, we can expect that the head noun of the NP (if any), will have the same set of features; subsequently, the rules dealing with accusative forms of singular feminine nouns get the highest degree of preference in further activation of the morphological component.

5) lexically stored information about inflectional patterns and irregular forms: the base representations of lexical entries are provided with declension and conjugation numbers if they follow regular inflection patterns. Suppletive and irregular forms are represented in the base lexicon.

Argumentation for similar approaches can be found in some previous studies (see Hammond & Noonan 1988). Using the numbers of the inflectional patterns, we express the assumption that the knowledge about them is in some way connected to the base lexical entries – as this knowledge must often be learned in the process of language acquisition. A learner of Swedish has to memorize the fact that the noun *katt* 'cat' has the plural form *katter*, while the phonologically similar noun *hatt* 'hat' requires the plural suffix *-ar*. A learner of Russian has to know which nouns belong to the subtype of 1st declension involving vowel deletion (as *zamok* 'lock', gen. *zamka*, cf. *vostok* 'east', gen. *vostoka*), while he/she must also learn the fact that the noun *put'* 'journey' has the inherent gender value 'male', but follows – unlike most male substantives – the pattern of the 3rd declension, etc.

The implementation of some parts of the bidirectional morphological model is illustrated by some examples of Prolog rules which may be used for analysis and generation of Russian noun phrases below.

In formulating the rules, we integrate the distinction between case function (CaseF) and overt case markers (OvCase) (cf. Bílý & Pettersson 1988).

A Russian subjective NP (rnps) in nominative is generated by the following RG-rule:

```
1. rnps(R,np(Rnux,Relcl)) --> rnux(R,Rnux,nom),rrelcl(R,Relcl).
```

An RG rule for object NP having the accusative case function is:

```
2. rnpo(R,np(Rnux,Relcl)) --> rnux(R,Rnux,acc),rrelcl(R,Reicl).
```


The entity 'rnux' contains the meaning codes of the head noun and its attributes except relative clauses (rrelcl). The constants 'nom' and 'acc' denote case functions. R symbolizes the description of the head noun (also called 'referent nucleus') which has the following shape in the Russian program version (leaving out the referent number in the first slot):

```
R= r(_,Meaning,_,Number,_,_,Animateness,Gender).
```

The empty slots are used in the machine translation process for inserting those grammatical values which are relevant for the equivalent noun in the other language.

As Russian attributes following the adjectival inflection pattern must agree with the head noun in case, number and gender, both the entity R, containing the relevant values, and the symbol of the case function must occur in rules describing the unit 'rnux' and its constituents. One possible variant of 'rnux' is:

```
3. rnux(R,h(Noun,_,Tot,Dem,Poss,Adj,_),CaseF) -->
   rtot(R,Tot,CaseF),          (totality marker, as vse 'all')
   rdem(R,Dem,CaseF),          (demonstrative pronoun, as èti 'these')
   rposs(R,Poss,CaseF),        (possessive pronoun, e.g. tvoi 'your')
   rap(R,Adj,CaseF),           (adjective attribute, e.g. krasivye 'beautiful')
   rn(R,Noun,CaseF).           (the head noun, e.g. dočki 'daughters')
```

Rules defining the constituents of 'rnux' may have direct access to the morphological component. At this stage, i.e. in the rules immediately activating the interplay between syntax and morphology, the distinction between case function and overt case marking becomes visible.

In Russian, only the singular nouns of the 2nd declension have overt accusative forms. In all other cases, the genitive (of animate nouns) or the nominative (of inanimate nouns) is used in the accusative function (cf. Bílý & Pettersson 1988). This can be expressed in the following way:

```
4. rn(r(_,Mean,_,Nb,_,_,Ani,G),Mean,CaseF) --> word(X),
   {caseform(X,r(_,Mean,_,Nb,_,_,Ani,G),Mean,CaseF,OvCase)}.
```

The formula can be read as: the syntactic constituent 'rn' (single noun) with a certain meaning and certain morphosyntactic values (Nb,Ani,G), fulfilling

the case function CaseF, may be realized as a word form X, if X is a caseform of a basic lexical entry having the same Mean(ing), Nb, Ani and G values and if this caseform (OvCase=overt case) may be – according to the current rules – used in the function CaseF.

The recognition and production of Russian genitive and accusative forms are illustrated below:

5. Accusative:

```
caseform(X,r(_,Mean,_,sg,_,_,Ani,G),Mean,acc,OvCase):-
        rlex0(Base,Mean,n,_,sg,Ani,G,nom,d2,_),
        suff(Stem,"a",Base),
        suff(Stem,"u",X),
        OvCase=acc.
```

6. The rule allowing the genitive of animate male nouns in the function 'accusative':

```
caseform(X,r(_,Mean,_,N,_,_,ani,ma),Mean,acc,OvCase):-
        caseform(X,r(_,Mean,_,N,_,_,ani,ma),Mean,gen,gen),
        OvCase=gen.
```

7. Genitive, 1st declension, 'strong' type:

```
caseform(X,r(_,Mean,_,sg,_,_,Ani,G),Mean,gen,gen):-
        rlex0(Base,Mean,n,_,sg,Ani,G,nom,d1,_),
        last(Base,Last),        (the nominative form does not end in a
        Last/=" ' " ,           palatalized consonant)
        suff(Base,"a",X).
```

Given a simple input sentence, like *mal'čik uvidel devočku* 'the/a boy saw the/a girl', the program finds the word form *mal'čik* (nominative) in the base lexicon, then identifies the form *uvidel* as a transitive verb and, subsequently, checks if the next constituent is an object NP (this means, that the rules for noun phrases with the case function 'acc' will be activated first). Then, the rule for 'rn' (rule 4) starts working. The meaning of the noun and its gender, animateness and number values are still unknown. The only constant values are 'acc' and the form which shall be recognized – here, *devočku*. At this stage, rule 4 looks like:


```
4a. rn(r(_,_,_,_,_,_,_,_,_),_,acc) --> word(devočku),
        {caseform(devočku,r(_,_,_,_,_,_,_,_,_),acc,_)}.
```

The next step is proceeding to the first rule for the accusative function, i.e. rule 5, which deals with overt accusative marker. By means of 5, the input form is successively matched with the stems of basic lexical items belonging to the 2nd declension ending in *a*. Stem matching is accomplished by the predicate 'suff' (based on the system-predicate 'append'), which may be used both for addition and subtraction of form endings. For example, given a command like: 'suff(X,"u",devočku)', the program answers: X=devočk, i.e. finds the stem, while the input command 'suff(devočk,"u",X)' results in producing the accusative form *devočku*. When the appropriate stem has been found, the program checks if providing the stem with the overt accusative marker *u* results in a string identical with the input form. If that succeeds, the meaning code and the number, gender and animateness values of the basic entry – here taken from:

```
rlex0(devočka,m(girl,sg),n,_,sg,ani,fe,nom,d2,_)
```

are transferred to the entity R (head noun description or 'referent nucleus') which receives the following shape:

```
r(_,m(girl,sg),_,sg,_,_,ani,fe).
```

After some subsequent stages (rules 4-2), the 'referent nucleus' is inserted in the representation of the whole object NP, and then in the functional representation of the whole sentence. When recognizing a form like *mal'čika* 'boy'+gen in the accusative function, we need a slightly longer procedure – as the rule for overt accusative (5) will fail, we have to find the appropriate rule for accusative function (6) and proceed to the rule producing and recognizing genitive forms of male nouns (7).

As mentioned above, the predicate 'suff' can be used bidirectionally – both for recognizing stems and adding endings. Therefore, rules 5-7 can be used for form production. When regenerating Russian sentences, or when translating from other languages into Russian, we start from a kind of scheme of utterance, containing syntactic roles and meaning codes of the constituents. The meaning of the word form which is to be generated is therefore already instantiated. Rule 4 has, subsequently, the following shape when used for

form-generation (e.g. in translation of the sentence *the boy saw the girl* into Russian):

```
4b. rn(r(_,m(girl,sg),def,sg,_,_,_,_),m(girl,sg),acc) --> word(X),
        {caseform(X,r(_,m(girl,sg),def,sg,_,_,_,_),acc,_)}.
```

The constants 'm(girl,sg)', 'def' and 'sg' are results of the input-analysis; the value 'def' obviously has no relevance for form generation, as the category definiteness is not used in Russian. The value 'acc' has been inserted by Russian syntactic rules after recognition of the transitive verb. The unknown value is in this case the form X, which will be produced by rule 5 by finding the basic entry with the meaning-code 'm(girl,sg)', i.e. the entry rlex0(devočka,...), checking the number, gender and declension symbols of this entry, finding the inflectional stem using the command 'suff(Stem,"a",devočka)' and adding the accusative marker *u* – again using 'suff'. The procedure is governed primarily by semantic and syntactic factors – only the lexical entries with meaning codes matching the input codes are involved, and the rules for the appropriate functional case (here accusative) are activated first.

For recognition and generation of adjectival inflectional forms, we use a similar procedure, but the particular rules dealing with adjectival patterns ('aform-rules') must be formulated in a slightly different way. The two main differences are the following:

– The base representation of adjectival entries are not their nominative forms, but stems, as there is no reason for choosing nominative forms of a certain gender as core representations: adjectives have no inherent gender.

– The adjectival inflection pattern may apply on stems subclassified as belonging to other categories – e.g. relative and demonstrative pronouns; thus, the aform-rules contain a special slot for the word-class symbol.

An aform-formula may look like 8 (a rule creating and recognizing genitive non-female singular forms belonging to the 'strong' pattern – i.e. the forms ending in *-ogo* in contrast to the 'weak' ending *-ego*):

```
8. aform(Category,X,r(_,_,sg,_,_,Ani,G),Mean,gen,gen):– G/==fe,
        rstem(Stem,Mean,Category,spattern,_,_,_,_,_,_),
        suff(Stem,"ogo",X).
```


Rule 8 allows correct identification and production not only in the case of typical adjectives, like *belogo* 'white'+gen, but also in the case of relative pronouns (*kotorogo* 'which'+gen), demonstrative pronouns (*ètogo* – 'this'+gen) and some other categories, among them nominalized adjectives. The last mentioned category is worth some further consideration.

In Russian, there is a group of lexemes having forms which follow the adjectival inflection pattern as far as their number and case are concerned. Their gender values are, however, inherent – as are the gender values of typical nouns. For example, the inflectional forms of the lexical entry *životnoe* 'animal' – genitive *životnogo*, dative *životnomu* etc. – have exactly the same endings as neutral adjectives like *beloe* 'white'+ne+sg+nom, *belogo* 'white'+gen, *belomu* 'white' +dat. But there is no masculine form *životnyj* meaning 'male animal' (there is an adjective with gender forms *životnyj, životnaja, životnoe,* but its semantic and syntactic properties are different – its meaning is 'having to do with animals' or 'bestial'). Forms like *životnoe* 'animal' function as noun phrase heads, giving the gender value to possible adjectival attributes, and – like typical nouns – they preserve their gender values when used as predicative complements, while predicative adjectives take the gender value of the subject:

sobaka – životnoe	'the dog is an animal'
dog+fe animal+ne	
sobaka bol'naja	'the dog is sick'
dog+fe sick+fe	

In the model suggested here, we assume that the presence of an inherent gender value is a sufficient criterion for treating a lexical entry as a noun. Still, we need the possibility of applying the aform-rules on such entries (which may be called 'lexically nominalized adjectives'). This can be accomplished as follows:

```
9. caseform(X,r(_,Mean,_,Nb,_,_,Ani,G),Mean,CaseF,OvCase):-
        aform(lna,X,r(Mean,_,Nb,_,_,Ani,G),Mean,CaseF,OvCase).
```

Rule 9 is to be interpreted as follows: Caseforms of a noun may follow the adjectival inflectional pattern; the subcategory of such nouns is symbolized by the constant 'lna' – lexically nominalized adjective, and the meaning of the

head noun in the 'referent nucleus' is identical with the meaning of the 'aform'. Then, the aform-rule applies on items subcategorized as 'lna' and provided with inherent gender values.

The treatment of adjectives which may function both as noun phrase heads (and having male or female forms in this function) or as attributes taking the gender values of the head noun is more problematic. Should such forms as *bol'noj/bol'naja* 'sick'+ma/fe be analyzed as nouns in the first case, but as adjectives in the second one (cf. *den sjuke/den sjuka* in Swedish)? Or shall we treat them as adjectives (when they occur as noun phrase heads, we may call them 'functionally nominalized adjectives')? Both approaches are possible, and this fact shows that the limits between derivational and inflectional processes are not completely clear.

The Prolog implementation of the approach suggested here is still at an experimental stage. The current version of the program runs quite slowly compared with the more practically oriented SWETRA programs. The speed can obviously not be seen as the most important criterion of a model's relevance, although it probably correlates roughly with simplicity. Both the speed and the linguistic relevance would increase if the program could recognize phonological similarities between word forms in a more efficient way. Implementing phonological similarities in Prolog is not impossible, but it is quite a complicated task.

2. MORPHOLOGICAL PROCESSING IN THE TRANSLATION PROCESS OF SWETRA

The translation process in the SWETRA machine translation system, like many other MT systems, consists of the three basic steps of analysis, transfer, and generation. The analysis step parses a source language input unit (sentence) and produces a functional representation (f-representation) of it according to the RG format. The transfer step transforms this f-representation into another, corresponding f-representation that is appropriate to the target language. The final generation step uses this new f-representation to produce target language output.

So far, relatively little work has been done within the project on the transfer step. In many cases it can still be defined simply as the identity transformation. (But see Gawrońska-Werngren 1989 for some ideas about how the transfer step can be used to handle definiteness in Russian-Swedish translation.) Instead, some care has been taken to design the f-representations

in such a way that they can be regarded as pseudo-universal, thus constituting an interlingua for the translation.

Morphological analysis inputs word forms and produces grammatical descriptions of them, including some representation of their meaning. In SWETRA, the grammatical descriptions can be coded as (partial) f-representations. Morphological generation inputs grammatical descriptions of words, including some representation of their stems or meanings, and outputs concrete word forms.

Morphological processing for translation always presupposes some kind of lexicon as background information. The output of such processing may either be directly inserted into the current representation of the text to be translated, or, in some cases, saved as entries in special, temporary lexicons to be used later in the translation process.

Morphological processing can take place at various points in the translation process sketched above. The main design choice concerns whether morphological processing should be integrated with the syntactic processing in the three basic steps or be done in separate processing steps intercalated between them.

2.1 Processing for analysis

The syntactic analysis step inputs a source unit and outputs an f-representation of it. Morphological information is needed during the analysis in order to determine whether a given source word form corresponds to a given grammatical description, and, if so, to retrieve the lexical information associated with the word for later use in the translation process. The morphological information can either be found ready-made in a comprehensive word form lexicon (lexical listing; see below) or be derived 'on the fly' by morphological rules using a core lexicon, which contains only base forms and irregular forms (rule integration).

Lexical listing

With this approach, all relevant word forms must have been collected into a lexicon in a pre-analysis step so that they are accessible to the syntactic analysis routine. The core lexicon is expanded in one way or other so that it includes the word forms found in the input. We may distinguish different subtypes of expansion, depending on what factors control it. In any case, a preliminary analysis of the word forms in the input is used to determine what base forms should be expanded. There is a trade-off between the depth of this

preliminary analysis and the number of (possibly superfluous) word forms that are generated. If the preliminary analysis is omitted altogether, the core lexicon must be fully expanded, so that the resulting extended lexicon contains all word forms in the source language (total expansion; see below). Otherwise only some part of the core lexicon needed for the analysis of the current source language unit is expanded (partial expansion; see below). The other extreme is a complete preliminary analysis, so that only entries for the word forms found in the input are added to the lexicon (direct word form identification; see below).

Expansion for analysis can be combined with expansion for generation. As source language lexical entries are created, target language entries can be generated for all meanings corresponding to the source forms. But if the target language needs more words, these have to be added as a result of transfer.

Total expansion

This approach generates all word forms and builds an extended lexicon. Consequently, it presupposes that morphological generation rules have been defined in order to generate all inflections of the base forms in the core lexicon. Since total expansion is a time-consuming process, it should preferably be done once and for all, prior to the translation of the whole text. A practical disadvantage of this method is that the extended lexicon will take up a large amount of space, which is a problem even with the memory facilities of modern computers. Search in such an enormous lexicon can be rather slow, unless the lexicon is stored in a particularly efficient, nonstandard manner. This may also make it necessary to produce separate versions of the lexicon for analysis and generation.

To illustrate how a lexicon can be morphologically expanded, we will first present a simplified rule of expansion and then some more detailed rules which may be used in the SWETRA modules.

It is natural to derive genitive forms in Swedish from the nominative forms by adding the (allo)morph *s*. The form *pojkes* is traditionally considered the genitive of *pojke, pojkens* is the genitive form of *pojken,* and *pojkarnas* the genitive of *pojkarna* etc. This goal can be achieved by writing the following Prolog rule:

```
genexpand :- slex(X,B,n,C,D,E,F,nom,H,I), suff(X,"s",Y),
             assert(slex(Y,B,n,C,D,E,F,gen,H,I)).
```


This formula means that if we write 'genexpand' the program will look for a lexical entry (X) with the word class 'n' and the case 'nom', add the genitive ending *s* to it making a new form (Y), and add (assert) this new form in a new lexical item with the case marking 'gen' but add all the other features unchanged to the (expanded) lexicon. If used recursively, this rule will add all possible genitive forms to the lexicon.

A little consideration indicates that this is not the whole truth. If nouns end in *s* they should not get another *s*, and we must therefore add a condition stating that the last letter of X must not be *s*. We may write this as: last1\=s (the last letter should be different from *s*). (The real rules must be even more complicated, as we will show in the conclusions).

It is natural to generate the genitive of the definite plural (*pojkarnas*) from the definite plural (*pojkarna)* and the definite plural (*pojkarna)* from the indefinite plural (*pojkar*). Such considerations make it natural to order the Swedish expansion rules in the following way: plur, def, gen. This means that all plural forms are first added to the lexicon. The definite forms are then generated from this expanded lexicon. In the last step genitive forms of all these forms are derived.

The genitive is a simple case – although not as trivial as shown. The addition of definite and plural morphs is a very complicated matter, where both phonetic (graphic) and morphological factors play a role. Generally, morphological rules can be simple if only phonetic (graphic) features in the form must be taken into account (as with the genitive). The rules can also be simple, if one can derive the inflection from the semantic class (or features) of the word. In fortunate cases, e.g. when all nouns denoting animals have the same inflection, and all nouns denoting trees belong to the same category, we do not have to define arbitrary morphological classes. But very often there is not such a clear relationship and the nouns must be grouped somewhat arbitrarily in morphological classes, traditionally called declensions for nouns and conjugations for verbs. To illustrate, there are two main gender classes in Swedish called 'reale' (re) and 'neuter' (ne), respectively. One cannot decide which class a noun belongs to on the basis of the phonetic (graphic) shape, the meaning (semantic features) or the morphological structure, and the gender classes must be considered as arbitrary today – although they may reflect some semantic classification in our ancestors' language. Similarly, the French division of the nouns between male (*le*) and feminine (*la*) genders is arbitrary for most nouns today.

The form of an inflected word is often the result of the interplay between morphological and phonetic (graphic) factors as is illustrated at several places in this paper.

Some unidirectional expansion rules for Swedish

The basic rule generating plural (plexpand) according to the first declension in Swedish could be formulated as follows (*kvinnor* 'women' from *kvinna* 'woman').

```
plexpand :- lex([A],m(Mean,sg),n,indef,sg,re,_,_,d1,F),
            suff(Stam,"a",A),
            suff(Stam,"or",Q),
            assert(lex([Q],m(Mean,pl),n,indef,pl,re,_,_,d1,F).
```

The rule states that a new lexical item with plural instead of singular is inserted into the lexicon if there is a certain singular word form (A) belonging to the first declension (d1) ending in *a*. The vowel is then deleted (using the procedure 'suff'), and the ending *or* is added to the stem.

The addition of definite forms of certain plural words of the fifth declension (d5) ending in 'are' is shown by the following rule.

```
defexpand :- lex([A],Mean,n,indef,pl,G,_,C,d5,F),
             suff(Root,"are",A),
             suff(Root,"arna",Q),
             assert(lex([Q],Mean,n,def,pl,G,_,C,d5,F).
```

This rule states that if there is a lexical item (derived before) which is indefinite, plural, belongs to the fifth declination, and ends in *are*, its root will get the ending *arna* and the resulting lexical item will be inserted into the expanded lexicon.

Expansion rules reflect the morphological competence of the language users in a special way. Such rules have been developed for substantial fragments of Swedish and Russian word forms.

Partial expansion

This approach expands only part of the lexicon. The aim is to generate only those word forms that occur in the current input unit to be translated. This is done once for each input unit, prior to syntactic analysis. The output should

preferably be stored as a small, temporary lexicon accessible to the syntactic analysis routine. Search in such small lexicons can be quite fast in comparison with search in complete lexicons. Another practical advantage of this approach is that lexical lacunae can be detected at an early stage in the translation process, before they would cause the analysis to fail, and typically after much useless backtracking.

As an example we mention a system to be used for analysis where the lexicon is expanded by all the forms beginning in the same three letters as the word which the system tries to identify. If e.g. the word *bilens* is looked for, but not found in the core lexicon, all words beginning in its first three letters *(bil)* in the core lexicon (e.g. *bil, bila, bilabial, bilaga, bilateral, bild, bilda, bilist, billig, biljard, biljett*) are run through the morphological expansion rules. This approach would generate a temporary lexicon including *e.g. bilar, bilen, bilarna, bils, bilars, bilens, bilarnas, bila, bilor, bilan, bilorna, bilas, bilors, bilornas, bilar* (verb), *bilade, bilat, bilas, bilas, bilades, bilats, bilabial, bilabialer, bilabialen, bilabialerna, bilabialens, bilabialernas, bilagor,* etc. Only one of these words is necessary (and most seem far-fetched), but we would find the form desired: *bilens*. Furthermore, we do not have to expand the whole lexicon and search all the resulting forms in order to identify the word *bilens*.

Empirical studies are needed in order to see how well such a system would work for different languages. It seems reasonable for Swedish, where the initial letters (sounds) are fairly stable over the word forms, and the first three letters seem to catch a characteristic part of the words, often the root, but sometimes a prefix. It is probably less effective for Russian, where the first three letters are not such a good characteristic.

For a language such as Arabic, one may use the traditional analysis based on three-consonantal roots and look for the three consonants of the word to be identified instead and establish a temporary lexicon by inserting all possible vowels between the consonants. It should then be possible to identify the word among these items. The advantage of this approach is, of course, that one must not have the whole lexicon available all the time in the processing of the sentence.

Direct word form identification

With this approach the word forms are identified by applying a series of rules that hopefully result in a temporary lexicon. Thus it is quite similar to the integrated approach, but outputs lexical entries rather than partial f-

representations. Another difference is that direct word form identification works with only one word form at a time, disregarding the wider syntactic context.

The classical morphological procedure is the stem-ending analysis. In this analysis the final part of the word is matched against a set of endings while at the same time identifying the stem in the (stem) lexicon. An experimental stem-ending analysis system for Russian has been investigated (by Gawrońska-Werngren).

Word forms may also be identified using the kind of bidirectional pattern matching rules worked out by Eeg-Olofsson 1988. Such rules can be used both in analysis and generation, and as they have been worked out in detail and work well for Swedish adjectives this approach will probably be used as the standard morphological procedure in SWETRA.

The basic idea is that morphology can be seen as analogies between concrete word forms and that the analogies are reversible (or bidirectional). One word form presupposes another, and that word form presupposes the first. More formally the analogies take the following form:

```
analogy(P0,G0,P1,G1),
```

where P0 and P1 are string patterns and G0 and G1 are the related semantic and grammatical (lexical) descriptions. Expressed in analogical terms: The relation between P0 and P1 is as the relation between G0 and G1. Such rules can be used both to derive the form P1 from P0 and to derive P0 from P1. The patterns P0 and P1 are given as strings, where certain elements are unspecified, while others are described by class membership or as constants. The analogy describing the relation between a positive adjective such as *vacker* 'beautiful' and its comparative form *vackrare* 'more beautiful' with 'unstable *e*' before a member of the class of the consonants 'lrn' is written as follows (simplified):

```
analogy([X,"e",lrn(L)],[...pos...],
        [X,lrn(L),"are"],[...comp...]).
```

We refer the reader to the paper mentioned for further details.

Morphological analysis by rules integrated with syntax

Typically, morphological analysis rules integrated with syntax analysis produce partial f-representations as output, but they may also output lexical entries, to speed up subsequent analyses of the word form in question. An advantage of the integrated approach is that it can be used to handle linguistic phenomena straddling the border between morphology and syntax (as shown above in section 1). A case in point might be Swedish particle verbs, e.g. *sätta in* 'put in'. Such verbs should presumably be represented as units in the lexicon. In a sentence, however, the particle and the verb proper may occur both solidly, written together in one word (*insätta* 'input', *insättning*), in sequence (*Han satte in sin hand* 'He put in his hand'), or separated syntactically, e.g. by an intervening adverb (*Han satte inte in sin hand*, literally: 'He put not in his hand'). From a theoretical point of view, integration seems to be a more satisfactory solution than separate morphological processing, since it obviates the problem of upholding a universally valid distinction between morphology and syntax. From the point of view of system design, however, integration is less desirable than a modular approach. Systems with complicated interactions can be very hard to understand and modify for practical use.

2.2.Processing for generation

The syntactic generation step inputs an f-representation and outputs target language word forms. Morphological information is needed during the generation in order to produce target language word forms corresponding to given grammatical descriptions (including some stems or meanings). As in analysis, the morphological information can either be found ready-made in a word form lexicon or be derived by integrated rules, using a core lexicon.

Lexical listing for generation

Lexical listing for generation aims at producing a lexicon that contains all the word forms needed by the generation step in the translation process. This can be achieved either by total expansion of a target core lexicon or by partial expansion. Total expansion for generation has the same disadvantages as total expansion for analysis (see above).

Partial expansion may be controlled by the meanings found in the target f-representations. All base forms with a particular meaning can be extracted from the target core lexicon, and morphological expansion rules may be applied to them to produce additional word forms to be stored in a temporary

target lexicon. Partial expansion should be carried out in a pre-generation step after transfer, because it is only after the transfer step that all relevant meanings are present in the f-representation. For instance, in Russian-Swedish translation the meaning corresponding to the Swedish copula *är* 'be' would generally not be present in the f-representation until after transfer (Russian has predicative sentences without copula).

Morphological generation by rules integrated with syntax

Integrated rules for generation offer the same advantages and disadvantages as integrated rules for analysis (see above). We will only add that it is particularly valuable if the rules can be made bidirectional, so that the same representation of morphological knowledge can be used for both analysis and generation.

CONCLUSIONS AND SOME RESIDUAL PROBLEMS

The previous presentations and discussions can be summarized as follows: It is possible to formulate and integrate morphological rules in several ways in Referent Grammar and in the translation system SWETRA.

1. Integrating morphological rules seems theoretically more interesting than identifying the word forms separately first. But there are clear programming and computing advantages in identifying the word forms as a first step in building a temporary lexicon. The syntactic rules then have only these few forms to work on, i.e. to find a suitable syntactic structure for.

2. There are several methods to use in the form identification process. Full expansion of the core lexicon with later search for the form to be identified is a clumsy method for languages with rich morphology such as Swedish – not to mention languages such as Russian and Georgian. Partial expansion controlled by a characteristic part of the target word (e.g. three letters) supplies a smaller temporary lexicon to search. It may be reasonable for Swedish, but further empirical studies are needed in order to evaluate this approach.

3. Form identification by stem-ending analysis is a good old method, but the pattern matching bidirectional rules based on analogy suggested by Eeg-Olofsson seem to be an interesting alternative.

4. Most morphological Prolog rules do not utilize all the phonological similarities and regularities formulated by phonologists in their rules, which are often based on distinctive phonological features. We will consider ways of remedying this situation.

5. In systems such as those discussed one may choose to call on interaction whenever a word form can not be identified. The user is then asked to supply the information, filling in the slots in the lexical representation. Another approach is to have the system guess on the basis of the phonological shape of the word form, the syntactic probability, and the frequency of the morphological patterns, etc. How such intelligent guesses should be implemented is an interesting problem. Clearly many words introduced by a capital letter can be assumed to be new proper nouns, but further empirical studies are needed to shed light on the whole problem.

6. The generation of Swedish genitive forms by adding the suffix *s* was discussed above. Further consideration indicates that there is more to the genitive than what was mentioned there. This can be exemplified by the famous phrase *Konungen av Danmarks bröstkarameller* (a trade mark; literally: 'The King of Denmark's sweets'). The problem here is that it is not the head noun of the noun phrase (konungen), which shows the *-s* but the last word in the phrase, a noun governed by a preposition. Grammarians (and academicians; cf. Wellander 1948) have noted the peculiarity and proposed that the correct wording should be: *Konungens av Danmark bröstkarameller* but this placement of the genitive *s* is felt to be very awkward by most Swedes.

Although only children or careless adults (cf. Wellander 1948) would say things such as: *Den flicka som kom igårs hund* (literally: 'The girl who came yesterday's dog', where the adverb *igår* seems to be inflected for genitive), *De som jag känners födelsedagar* (literally: 'Those who I know's birthdays', where the verb *känner* has got the *s*) or *De som har kommits ansökningar* (literally: 'Those who have come's applications', where the participle *kommit* gets the *s*), there are enough cases indicating that the genitivization rule could operate on the whole np rather than the head noun and that whatever word is last in the np tends to receive the *s* unless the construction is avoided. This is certainly a complication for morphology, and we leave it to the reader to figure out how such rules should be written.

REFERENCES

Anderson, S. 1982. 'Where is morphology?' *Linguistic Inquiry* 13, 571-612.

Bílý, M. & T. Pettersson. 1988. 'Neutralization in case morphology'. *Working Papers* 33, 5-21. Dept. of Linguistics, Lund University.

Eeg-Olofsson, M. 1987. 'Software systems for computational morphology – An overview'. *Nordic Journal of Linguistics* 10, 1-34.

Eeg-Olofsson, M. 1988, 'A morphological Prolog system for Swedish based on analogies'. *Papers from the first Nordic Conference on Text Comprehension in Man and Machine*. Ö. Dahl, & K. Fraurud, eds. Dept. of Linguistics, Stockholm University.

Gawrońska-Werngren, B. 1988. 'A Referent Grammatical analysis of relative clauses in Polish'. *Studia Linguistica* 42, 18-48.

Gawrońska-Werngren, B. 1989. 'Identifiering av diskursreferenter vid maskinöversättning från ryska till svenska'. *Handl. från de nordiska datalingvistdagarna,* Iceland July 1989 (to appear).

Hammond, M. & M. Noonan. 1988. *Theoretical morphology*. San Diego: Academic Press.

Hellberg, S. 1978. *The morphology of present-day Swedish*. Stockholm: Almqvist & Wiksell

Karlsson, F. 1986. 'A paradigm-based morphological analyzer'. *Papers from the Fifth Scandinavian Conference on Computational Linguistics,* ed. F. Karlsson. Helsinki: Dept. of General Linguistics.

Matthews, P. 1947. *Morphology*. Cambridge: Cambridge Univ. Press.

Sigurd, B. 1987. 'Referent Grammar. A generalized phrase structure grammar with built-in referents'. *Studia Linguistica* 41, 115-35.

Sigurd, B. & B. Gawrońska-Werngren. 1988. 'The potential of SWETRA – A multilanguage MT system'. *Computers and Translation* 3, 237-50.

Wellander, E.1948. *Riktig svenska*. 3rd ed. Stockholm: Svenska Bokförlaget.

Lund University, Dept. of Linguistics
Working Papers 35 (1989), 191-202

Shanghai Vowels

Jan-Olof Svantesson

In this article I present acoustic data on Shanghai vowels, and make an analysis of the system of vowel phonemes based on these data.

THE VOWEL SYSTEM

My investigation is based on Sherard's 1980 description of the Shanghai vowel system. He describes Shanghai as having the following fourteen vowels:

i	y	ɿ	ʅ		u
ɪ	ʏ				
e		ɚ		ɤ	o
ɛ					ɔ
		a			

(The correspondence between my symbols and his is: ɪ=y, y=ü, ʏ=ö, ɿ=i̇, ʅ=ï, ɚ=ər, ɤ=ɔ̈.)

Not all of these are phonemes, however: from the table of Shanghai syllables given by Sherard, it can be seen that the apical vowels [ɿ] and [ʅ] occur only after dental/alveolar sibilants and affricates (his [s], [z], [ts], [ts^{h}]), while [i], [ɪ] and [y] never occur in this position. Thus, [ɿ] and [ʅ] can be regarded as allophones of the corresponding non-apical vowels /i/ and /y/.

The vowel [ɚ] may be regarded as a syllabic [ɹ̩]. Like the syllabic nasals [m̩], [n̩] and [ŋ̩], it has a restricted distribution, not combining with initial consonants, but always forming a syllable on its own.

There is a contrast between what are traditionally called 'clear' and 'muddy' syllables in Shanghai. This is a prosody which affects both the vowel and the initial consonant: in muddy syllables, the first part of the vowel has a special phonation sometimes referred to as voiced aspiration, and obstruent consonants have shorter duration in muddy than in clear syllables. Furthermore, initial sonorants are preglottalized in clear, but not in muddy syllables. Muddiness will be symbolized by ˍ written under the vowel. In