Designing and testing linguistic development phases in a machine translation project

Bente Maegaard,
Institute for applied and mathematical linguistics
University of Copenhagen
Njalsgade 96
DK-2300 Copenhagen S
Denmark

In the first sections of the paper the design of a scheme for building up linguistic coverage in a multi-lingual machine translation project (EUROTRA) is discussed and a solution is proposed which also takes into account the extensibility of the system, and in the last section the aspect of testing is briefly discussed.

1. The environment.

This paper concerns a concrete project, EUROTRA, but the problems described are of a general nature.

Those basic characteristics of EUROTRA, which are relevant here, are the following. EUROTRA aims at producing a pre-industrial prototype for machine translation between the 9 languages of the European Community. The vocabulary to be covered is around 20,000 lexical items within specific text types and within a specific subject field.

EUROTRA is divided into phases, each with their sub-goal. The goal of the second phase is the development of a small-scale translation system (all languages), for a limited vocabulary (2300 items), based on a corpus text. The goal of the third (and final) phase is the development of a more general system (not corpus based, extensible) with the coverage mentioned above.

NOTE: In this description we have only mentioned the goals of linguistic development, and deliberately left out all other aspects of the project (research, software etc.).

It should be mentioned as well, that the treatment of the Spanish and the Portuguese languages is a little delayed, because Spain and Portugal joined the Community after EUROTRA had started.

The project can thus be characterised by three important features:

1) medium-scale
2) extensible
3) multi-lingual

A fourth one is that it is corpus-based at one stage, but has to develop into a more general system.

The fact that the project is medium-sized means that it is necessary to define various development phases for the linguistic coverage - it cannot be built in one single block.

Furthermore such an approach whereby the full task is broken down into smaller tasks, supports the final goal that the system should be extensible, because the extensibility is built-in and is being tested already during the development.

The fact that the project is a multi-lingual ma-
possible, and secondly, the actual coverage of the
system becomes unknown.
Therefore we have made a description of what the
grammars of the various sub-periods should cover
(an intensional definition). It follows from the
discussion above that such a definition of the
coverage of grammar and vocabulary has to meet the
following conditions:
1) it has to describe equivalent phenomena in
the various languages
2) it should be possible to extend the
linguistic coverage, without throwing away
(too much of) the grammars and dictionaries
already produced.

Point 2) above led to suggesting that in the first
period one should concentrate on developing a sys-
tem containing the skeleton of a sentence and its
main building blocks. Therefore the first defini-
tion of linguistic coverage covers main clauses
with no dependents (with one exception). Such sen-
tences are of course quite simple, but it should be
noted that the main clauses may contain adverbials
in all possible positions, and all arguments in all
possible orders. The idea is that e.g. adverbial "slots"
in a next version of the grammar is then expanded to be filled not only by adverbs, but by
adverbial clauses.

An overview of the scheme made for the development
of the linguistic coverage in the second phase of
EUROTRA is given in the appendix. In the following
we will comment on the reasoning behind it.
At the constituent level we have first of all the
noun phrases. All types of noun phrases are treated
in the first round, i.e. all types of modification
of the noun itself or any of its modifiers includ-
ing participles. This inclusion of participles
entails the inclusion of relative clauses in order
to make translation possible, cf. translation from
French into Danish of the above examples
s'occupant, intéressées.

In the first round no control verbs are treated, as
this would add the complication of empty elements
and co-indexing. By the same rule no modal verbs
are treated in the first period. Auxiliary verbs
however are accepted, as they together with the
main verb form one unit at a later stage of analy-

sis.
The fact that modal verbs are excluded has led to
exclude also the future tense, - as the future
auxiliary in many languages is a modal verb. Also
other verb tenses the following are treated: only
indicative, both active and passive, present and
past tense, and tenses made by combination of pre-

tent and past tense of auxiliaries with participles
(imperfect, pluperfect). We do not include sub-

junctive, and not infinitives.

In order to avoid co-indexing, also some of the
pronouns have been omitted: personal and demonstra-
tive. These pronouns include are possessive, rela-
tive, reflexive, indefinite, and all adjectival
pronouns (not because these do not involve co-
indexing, but because missing co-indexing is hoped
to be less damaging in these cases).

In the second period the following complications
are added: Subordinate clauses, adverbial as well
as nominal. These clauses may take the place of
simple adverbials or noun phrases of the first
period. This means that the grammar rules speci-

fying sentence patterns has to be slightly modi-

fied.
Furthermore participle constructions with
sentential function are added. As mentioned above
participles modifying a noun were part of the first
period. Control verbs and infinitives are also
added, and simple coordination (coordination of
noun phrases, adjectival phrases, adverbs, preposi-
tional phrases).
Participles, control constructions and coordination
all require empty elements and co-indexing. Correct
treatment of relative and other pronouns also
involve the use of a co-indexing mechanism.

Finally, in the second period as well as the
first period, modal verbs and other
modal expressions are included, as well as various
types of movement phenomena. It has been foreseen
also to include treatment of parenthetical
insertions and appositions, and a better treatment
of pronominal reference, but this may have to be
postponed to the third phase of EUROTRA.

2a. Comments on levels of description
In the above short survey we have been using only
syntactic and morpho-syntactic criteria. But as any
other natural language project EUROTRA operates
with deeper levels of description as well.
The definition of the linguistic coverage has to
take phenomena at these levels into account as
well. Here we may take the verbal tense / verbal
time as an example.
In the first period only main clauses are treated
and only some tenses and only at the morpho-
syntactic level (awaiting a semantic legislation for
the representation of time). In the second period
the time legislation, i.e. the deep representation
of the tenses, is implemented, and in the third period
it is extended to subordinate clauses at the
surface as well as the deep levels.

This is of course one way of defining linguistic
coverage. It could be argued that a more reasonable
approach would be to start from the interface
structure definition which is common for all
languages and define the coverage in terms of this.

We find at least two arguments against this: first
of all the practical one that a full definition of
the interface structure was not ready when the
first implementation started. Secondly, the
linguistic data which have to be treated are ex-
pressed as surface text and it seems more reason-
able to define coverage systematically in terms of
this surface representation, than in terms of the
abstract representation.

2b. Lexical coverage
An aspect of linguistic coverage which has not been
treated above is the lexical coverage. The lexical
items are of course taken from the corpus of the
second phase of EUROTRA. But a definition of lexi-
cal coverage consists in more than just defining
the vocabulary: it also consists in defining the
content of the dictionary, the number of readings
to be distinguished, the feature system to be used.
The reason that the vocabulary and the number of
readings cannot be seen as being defined by the
corpus text itself is that this would be too speci-

fth and hence too unsystematic, i.e. not easily
extensible. Here the question of extensibility may
be a little different then for the grammar. Extens-
ability of a lexicon in terms of adding new items,
using the same features as in an earlier version of
the dictionary, presents no problem. But when the addition rather consists in adding new distinguishing features, i.e., new readings, all the relevant lexical entries have to be checked for modification. The only measures which can be taken to facilitate this type of extension of the dictionary is to use a reasonably well-structured set of features, so that extensions may concern only one feature or a few features at a time.

3. Conclusion remarks on extensibility.

While not claiming that the above defined scheme of progressively growing linguistic coverage is the only possible one, we believe to have shown that it is a reasonable one, with respect to the languages involved, and with respect to extensibility. The modifications of the grammar which are necessary when going from one period to the next can in most cases be made very locally. Take as an example the extension of a noun phrase to comprise sentential complements; this can be done almost solely by additions to the grammar, but obviously a few modifications of the existing grammar cannot be avoided.

Furthermore we want to add some comments on the possible definition of the linguistic coverage in the third phase of EUROTRA (and maybe beyond). It may well be that, taking into account the complexity of the system, and the multitude of languages, it will be more revealing to define the linguistic coverage in a negative way: by stating the phenomena which are not treated. Internally however, in the project, and in particular in the language groups, the explicit, intensional definition of coverage will always be needed, and will be the basis of the linguistic design.

Before we leave this section on definition of linguistic coverage we could add information on the actual status of implementation: the first period coverage was obtained for most languages during spring 1987, second period will be obtained early 1988, and third period mid 1988, for all the main modules: analysis, transfer and generation.

4. Testing

This is a very brief sketch of the types of testing needed to check the linguistic coverage. The type of testing which is adequate is of course dependent on the way in which the coverage has been defined:

If the extensional definition of a corpus has been adopted, then the testing is very simple: check if the corpus can be treated adequately.

If an intensional definition is adopted like the one suggested above, the question of testing becomes less simple, because the claims of the system are more general: all sentences described by the grammar and in particular in the language groups, the explicit, intensional definition of coverage will always be needed, and will be the basis of the linguistic design.

Appendix: schematic overview of the elements of the sub-periods.

Period 1:
- Sentences containing one main clause with a verb
  - in any tense except for morphologically expressed future
  - in active or passive
  - in indicative.
- Constituents that do not contain sentences, except for modifiers with verbal governor in a noun phrase (participles).
- Constituents that do not contain sentences, except for strictly subordinated, modifying relative clauses and participial constructions.
- Fully expanded noun phrases except for appositions (this includes all kinds of modification with adjectival phrases, numerals, prepositional phrases etc.).
- All pronouns except for personal and demonstrative pronouns.
- All adjectives, including pronominal adjectives.
- All adverbs.
- Coordination, only for simple noun phrases.

Period 2 (extensions):
- Subordinate clauses, adverbial and nominal.
- Infinitives governed by control verbs, and "free" infinitives (infinitives governed by modal verbs are excluded).
- Participles with sentential function.
- Sentential arguments for nouns and verbs.
- Coordination of noun phrases, adjectival phrases, adverbs, prepositional phrases, but excluding verb phrases and clauses.
- Verbless sentences.
- Time in main clauses

Period 3 (extensions):
- Modality.
- Time in subordinate clauses.
- Movement.
- Apposition, parenthetical insertion.
- Pronominal reference.

Acknowledgement
The definition of the linguistic coverage of the second phase of EUROTRA was made by Charlotte Toubro (then: EUROTRA-DK, now: ALPS, Switzerland) and the present author.

We also want to thank the many collaborators in EUROTRA who consented on this definition, thereby improving it.

Reference
Alan Helby (no title), Proceedings of COLING86, p.104-106.