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 (2500 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 machine translation project implies that for each language the same or equivalent phenomena should be treated at the same time, so that translation is possible.

Below we will describe the development of the linguistic coverage in the second phase of the project.

In the second phase we are working with one single "corpus" text which exists in all the 9 languages (Danish, Dutch, English, French, German, Greek, Italian, Portuguese, Spanish). In fact the goal of the second phase is to create a machine translation system which is able to translate this text. The text is the Commission of the European Communities' proposal for the ESPRIT programme, some sample texts are shown below, to give an idea of the similarities and the differences.

One can e.g. notice that participle constructions may correspond to relative clauses (in some languages, like English, there is a choice between the two, in others, like Danish, only one is possible), one can also notice the complex use of future and modality in will need, devront (avoir), skal (have), braucht. These are just typical and well-known examples of differences/similarities between the languages treated.

Danish
De, som er involveret i udvikling og anvendelse af programmel, skal have adgang til programmeludviklingsværktøjer i et gradvis mere integreret miljø, og de, som har med datamatstøttet konstruktion (CAD) ....

English
Those involved with the development and use of software will need access to software development tools, in a progressively more integrated environment, and those involved with computer aided design (CAD) ....

French
Les personnes s'occupant du développement et de l'utilisation du logiciel devront avoir accès aux outils de développement de plus intégré, et celles intéressées par la conception assistée par ordinateur (CAO) ....

German
Wer an der Entwicklung und dem Einsatz von Software beteiligt ist, braucht im Rahmen eines schrittweise stärker integrierten Umfeldes den Zugriff zu Software-Entwicklungswerkzeugen. Wer sich mit dem rechnergestützten Entwurf (CAD) ... befasst ...

2. The definition of linguistic coverage.

A very obvious way of ensuring correspondence between the phenomena treated in the various languages is of course to treat the first n (n = 5/10/25) pages in each language (extensional definition of coverage), because it is given that the texts are equivalent. But such a method would have nothing to do with designing, as one would get a random collection of linguistic phenomena. Furthermore the extension to a greater coverage (e.g. next 20 pages) can normally not be done in a systematic way (because the point of departure was not systematic). It is well-known that this working method has several disadvantages: it gets easily out of control, so that error correction becomes im-

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 system contaning the sceleton of a sentence and its main building blocks. Therefore the first definition of linguistic coverage covers main clauses with no dependents (with one exception). Such sentences 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 af 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 including 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 analysis.

The fact that modal verbs are excluded has led to exclude also the future tense, - as the future auxiliary is in many languages a modal verb. As for other verb tenses the following are treated: only indicative, both active and passive, present and past tense, and tenses made by combination of present and past tense of auxiliaries with participles (= perfect, pluperfect). We do not include subjunctive, and not infinitives.

In order to avoid co-indexing, also some of the pronouns have been omitted: personal and demonstrative. The pronouns included are: possessive, relative, 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 specifying sentence patterns has to be slightly modified.

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, adjective phrases, adverbs, prepositional 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.

Furthermore, as a complication to noun phrases and adjective phrases, sentential arguments for nouns and adjectives are added.

The last type of grammatical construction, which is added in the second period, is verbless sentences (headlines, titles etc.). As sentences they have a different grammar, but also as noun phrases they may have a slighlty different syntax.

Finally, in the third 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. Comment: 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 expressed as surface text, and it seems more reasonable 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 lexical 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 its number of readings cannot be seen as being defined by the corpus text itself is that this would be too specific and hence too unsystematic, i.e. not easily extensible. Here the question of extensibility may be a little different than for the grammar. Extensibility 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. Concluding 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 complement; 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 the lexicon should be treated adequately, and such a set of sentences will normally be infinite.

Here it seems reasonable to combine two approaches: first of all, the conduction of a systematic test, whereby all types of constituents, and a reasonable amount of combinations of constituents are tested. And secondly, also to test the grammars and dictionaries developed against "real" text. This last testing allows for random combinations that were not taken into account by those devising the systematic test. All this testing should be done after the implementors' testing is performed, and by a different group of people.

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 commented on this definition, thereby improving it.

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

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 are themselves not 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 participal 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, adjective 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.