

# LOGON

Workplan, 2005–2006

## 1 Introduction

The original work plan divides the project period into cycles. The first cycle, until March 2004, should serve as a first test period for the approach and a first demonstrator be made. The second period, until September 2004, should serve as a period for evaluation, reflection and redirection. The plans for the third period, Oct.2004–Dec.2006, were not very detailed. They should be decided in September 2004 on the experiences made so far.

The original project proposal also contains many loose ends and open questions which might or might not be pursued. After the first project period we have become richer in experiences and acquired a better background for making decisions. Some choices have been irrevocably made, while some of the open questions from the original proposal remain to be decided.

This document will make the plans for the remaining project period more concrete. It will be more detailed on the first year, 2005, than on the second year, 2006. Exactly what we will do in 2006 will depend on the results from 2005, and we will make a more detailed work plan before the start of 2006.

## 2 Where are we?

The start of the project was slower than anticipated in the work plan. This is a common experience in larger projects, often referred to as the S-curve. We started with three half-time researchers and the first doctoral stipendiary in Jan./Febr. 2003, got another full-time researcher from July 2003, while the last 3 doctoral stipendiaries started in Jan. and Febr. 2004. From April 2005 and for the rest of the project period, we will get one more full time researcher (Dan Flickinger).

Our main result so far is that we have made a first demonstrator (in fact several progressively better versions.). The demonstrator is based on MRS-transfer, where the generation is done in HPSG/LKB/ERG and the analysis is done in LFG/ XLE/NorGram extended with MRSs, as sketched in the original work plan. The demonstrator up to the Sept. 1-version (“Kaffebrenneriet”) was developed on the basis of a restricted test set of 104 sentences from the hiking-domain. The latest version of Dec. 1 (“Con gusto”) was in addition based on a hand-constructed test suite of closed-class items (of close to 300 sentences).

The demonstrator shows that the framework is feasible. It is possible to build the different modules together and use the result for translation. On the other hand, it also indicates that the task is harder than expected in some respects. The evaluation group has done an evaluation presented in a separate report which gives an impression of the performance of the Sept. 1, 2004 demonstrator (“Kaffebrenneriet”). The original work plan anticipated a 95% success rate on the training material (500 sentences) and an 80% success rate on unknown texts at the end of the project period. Right now, this does not seem to be within reach.

We can conclude from this first evaluation round to continue with the same basic architecture for the backbone of the demonstrator. The work plan opens for other representations than MRSs, but it is clear by now that we will stick to the MRSs (but possibly modify and extend them). We will also stick to our approaches to analysis (LFG extended with MRSs) and generation (HPSG/LKB/ERG).

The original work plan was also open-ended when it came to the themes for the doctoral projects leaving a certain freedom to the stipendiaries. The doctoral projects are

**Petter Haugereid** *Automatic translation via MRS representations*

**Gunn Inger Lyse** *Oversettelsesbasert semantisk informasjon  
for automatisk læring av flertydighet*

**Liv Ellingsen** *Myke velformethetsbetingelser*

**Erik Velldal** *Stochastic ambiguity management for MT*

### 3 Main challenges

We see four types of main challenges for the remaining project period: (i) To produce good research and get it properly published, (ii) to develop the functional demonstrator so it can serve as a proof of concept of our approach, (iii) to produce reusable resources for Norwegian language technology, (iv) to make sure that our stipendiaries complete their doctoral and masters projects. Let us expand a little on each point.

**Publications** Compared to the overall goals of the KUNSTI program, we are doing quite well on several parameters, like national and international collaboration, and the move towards technology and development of a demonstrator. But we are far behind on one point — publications. So far (Sept. 2004) we have invested 4 person years at the post-doctoral researcher level in addition to the (unpaid) PIs, but we have mainly published some system descriptions in conference proceedings. A main reason to this is that we have invested most of the mentioned resources in the development of the

demonstrator. The challenge now is to make sure that the work—also the work related to the demonstrator—at the same time results in papers which get published.

**Proof of concept demonstrator** We have shown that we can build a demonstrator based on our model. But this demonstrator has so far a very limited coverage. The challenge now is to show that the approach scales up beyond a toy system and at the same time has favorable properties compared to other approaches.

We can recognize several more specific subtasks here. First of all, the system has to be able to handle more sentences. It has to be extended to cover running text within a restricted domain. Second, we have to face the problems of ambiguity. The original work plan mentions ambiguity management, but puts it somewhere in the periphery. The work so far has shown that we cannot ignore this problem.

**Reusable resources** One goal of the KUNSTI program is to produce reusable resources for language technology on the Norwegian language. It is a goal that resources that we build as parts of our demonstrator also can be made available for other applications and, as far as possible, be made freely available. This should include a lexicon, a tokenizer, a morphological analyzer and a grammar for Norwegian.

**Building competence in Norway** Another stated goal for the KUNSTI program is to strengthen the competence in language technology in Norway—in particular by educating Ph.D. and master candidates. We have appointed four doctoral stipendiaries. It is a vital goal for the project that they complete their degrees. It is also a goal to produce good master candidates, and we will continue to use masters stipends for the remaining project period.

There is a potential conflict between these goals, in particular between developing the quality of the demonstrator and producing research publications. The overall attitude to this will be the following. This is a research project. The demonstrator shall be developed to be a genuine proof of concept. Thus, the overall goal cannot be stated in terms of the best possible coverage, precision or BLEU score. The full array of well-known techniques found in other MT efforts to enhance performance does not have to be developed and added as part of the demonstrator. They can of course be added in a later stage if the demonstrator is developed into a working system. But we shall not spend time and resources on them in the project unless such work can be expected to result in new research which can get published.

## 4 What shall we do?

### 4.1 Organization

In the original work plan, the project is divided into *subprojects*. Later the concepts of *component* and *component maintainer* were introduced. This was overlapping with the division into subprojects but did not reflect it directly. We have also used the term *work package*. It is not obvious that the original division into subprojects is fruitful anymore. First, the work on the demonstrator has shown that we need a tighter collaboration than what is indicated by the subproject model. Second, we need to break down the work into smaller packages than the original subprojects.

The demonstrator consists of components which have to be developed and maintained throughout the entire project period. Work packages may be part of the construction or development of a component. They may also cut across several components. In some cases the distinction between component, work package and deliverable is sharp; in other cases it is more blurred. The main point is that they are all assigned a responsible person.

### 4.2 Backbone

The project will, as before, be organized around the demonstrator with its three-stage symbolic backbone: analysis, transfer, generation.

**Component 1:** Analysis, system

**Responsible** Paul

**Component 2:** Analysis, grammar

**Responsible** Helge

**Component 3:** Transfer, system

**Responsible** Stephan

**Component 4:** Transfer, grammar

**Responsible** Jan Tore

**Component 5:** Generation, system

**Responsible** Stephan

**Component 6:** Generation, grammar

**Responsible** Dan

**Component 7:** System integration

**Responsible** Stephan

One change compared to the first period is that Dan will be working full time on the project, and he will be the natural maintainer of the generation grammar. The transfer grammar turns out to be a bigger task than originally anticipated. It will involve several people, hopefully at all three sites. It is natural that the integration and coordination of this is placed together with the responsibility of the project as a whole.

### 4.3 Common core tasks

In spite of the division into components and responsibilities, the work on the demonstrator shall be a joint effort with close collaboration. For the rest of the project period we will fix a reference text/corpus. We will continue to work on the hiking domain. The hiking domain is chosen because we already have invested some efforts in this domain, it has obviously an applied value, and it turns out to contain interesting linguistic constructions which are not often discussed within linguistics.

This text will decide our domain. The vocabulary should not be chosen too big (and of course not too small), as we are building a proof of concept demonstrator and not a real application:

- Around 5000 lexemes from Norwegian should suffice. We expect 3000 sentences with an average length of 15 words to be suitable.
- We shall get the text translated. We shall have two (three?) independent translations.
- All translators are instructed to provide a direct, but idiomatic, translation without paraphrases. One of the translators is specially instructed to provide a translation as close as possible to the source text.
- The text shall be without copyright restrictions.
- From the text, a part is kept aside (from the developers) as a test corpus. The test corpus is divided into two: a known-vocabulary, and an unknown-vocabulary corpus. The vocabulary of the first part is extracted and made known to the developers.
- The original and translated texts should be tagged and the source and translations should be sentence aligned.

In addition to this text, we should also make sure that the demonstrator covers a core vocabulary (say 1500) words and the main constructions from the source language .

**WP 8.** The reference corpus

**Responsible** Janne and Torbjørn

The analysis and generation grammars are presented as two different components. To reach our goals it is important that the development of the two are harmonized.

**WP 9.** Cross-lingual semantic harmonization

**Responsible** Helge and Dan

We plan building tables for semantic interfaces (SEM-Is) for the two languages serving as a form of interchange format between the transfer component and the two grammars. First, the format and technical specification of such a SEM-I has to be further developed; implementation of the SEM-Is will require some software support (SEM-I-internal consistency checking and MRS wellformedness testing). Then it can be filled with content.

**WP 10.** SEM-I

**Responsible** Dan and Stephan

#### 4.4 Theoretical studies related to the core architecture

In addition to the direct work on the back-bone demonstrator driven by the particular text corpus, i.e., a bottom up-approach, we may at the same time pursue a more traditional top-down approach. Here we single out some particular linguistic phenomena where we lack satisfactory analysis and where some efforts may result in

- better linguistic analysis related to each of the two languages
- interesting contrastive analysis, or
- better translational analysis.

This work relates to the point on experimenting with richer lexical representations mentioned in the original work plan. If this work results in analyses that enhance translation performance it might be implemented as part of the demonstrator. Irrespectively of that, this work will be theoretically motivated and result in publications. All sites are expected to contribute to individual studies, and sufficiently self-contained sub-phenomena may form the basis for master student projects. Possible phenomena include:

- Prepositions
- Indefinite NPs with and without articles
- Mass and count terms
- Modals
- Tense and aspect
- Degree specification
- Complex proper names

**WP 11.** Theoretical studies related to linguistic phenomena

**Responsible** Lars and Dorothee

The original work plan also mentioned a more theoretical study of underspecified semantic representations, like MRS, and their use as transfer formalisms. It might be worthwhile to return to this question in light of the experiences from the project, e.g. the analysis of degree specifiers.

**WP 12.** Theoretical studies of MRS formalism

**Responsible** Dan and Jan Tore

## 4.5 More detailed planning related to the back-bone

Many of the tasks related to the development of the demonstrator, in particular the extensions to the analysis and generation grammars, are sufficiently specified from the descriptions so far. But some of the other steps might profit from a more detailed planning.

### 4.5.1 Morphology and preprocessing

Currently, Paul is working on the integration of a morphological analyzer into the analysis component. This analyzer is based on the work on the OB tagger and will give us full control of the process. In particular it will add a compound analyzer and the possibility of name recognition. These are also the general language technology tools that LOGON should aim to prepare for re-use outside of the project, i.e. produce portable stand-alone versions and work on packaging, documentation, and distribution.

**WP 13.** Morphology and preprocessing

**Responsible** Paul

### 4.5.2 Transfer

The transfer formalism and component are less developed than analysis and generation and will need further work. We currently have a stable transfer formalism, but the transfer rules have a very limited coverage. We are not sure exactly what will be the best procedure to extend the component radically, but the following parts seem to be involved.

1. further development of the transfer formalism
2. establishing correspondence types, i.e. transfer templates
3. fill in the actual correspondence instances corresponding to the types
4. consider ways of acquiring the correspondencies (semi-)automatically

A specific task related to the further development of the transfer formalism is the possibility of MRS-packing, i.e. enriching both the MRS and transfer rule formalisms to allow a compact representation of multiple hypotheses in

a single MRS. This could take several forms, for example the introduction of a notion of optionality at the EP level, atomic disjunction of semantic predicates, or a full disjunctive encoding. Assuming MRS packing was available, transfer could avoid the explosion in combinatorics introduced by alternate (lexical) transfer rules; likewise, multiple analysis results could be packed into a single transfer input, and the generator would have to support generation from packed MRSs. This facility raises an interesting research question, viz. the interaction with stochastic ranking at each processing level, where presumably it is *not* desirable to just accumulate disjunctions throughout the translation pipeline, but rather strike a good balance between pruning highly unlikely hypotheses and maintaining enough of a beam to avoid ‘early commitment’ errors.

**WP 14.** MRS-packing

**Responsible** Stephan

As an input to the construction of the transfer rule instances, it will be very useful to get a survey of the possible (lexical) translation correspondences. We may extract the Norwegian and English vocabulary from the example texts. From this we can align the possible translations of the Norwegian words and try to classify the conditions under which these translations are possible. This task will have strong connections to the SEM-Is (above) and the lexicon (below), in addition to the transfer rules.

**WP 15.** Survey of translation correspondencies

**Responsible** Victoria

Ways to (semi-)automatically extend the transfer grammar (or transfer lexicon) from examples shall be investigated.

**WP 16.** Ways of (semi-)automatically extending the transfer lexicon

**Responsible** Torbjørn

### 4.5.3 Improvements to the generator

One particular improvement to the generator (maintainer Stephan) to be pursued is the task of simplifying the generator input, i.e. support for additional underspecification and automated, generator-internal insertion of ‘predictable’ pieces that are required for generator success. Known examples include particles and selected-for prepositions (as, say, in dative shift, passivization, partitives, and possessives). But we should take a somewhat wider perspective and investigate the nature of parts of MRSs more generally. With only minimal assumptions about the generation grammar, elements like messages and handle constraints or quantifiers under certain conditions, should be predictable. Thus, they need not (always) be supplied in the generator input.

Another desirable property is generator robustness, similar to fragmented analyses in parsing. Even though an MRS does not correspond to a full sentence, could it still be possible to generate meaningful fragments from it?

## 4.6 Ambiguity management

The original work plan recognized ambiguity as one of the largest challenges for an MT system, while on the other hand placed the problem somewhere in the periphery of the project. The experience from the first project period is that we can get as much as 14,000 different outputs from one sentence. We cannot ignore this problem in the final demonstrator.

We need a project-wide discussion of the goal of the demonstrator with respect to ambiguity. First, if a Norwegian sentence has several different possible analyses, is the goal to translate all these different analyses, or only the most likely analysis given the textual context where the sentence occurred, or only the most likely analysis irrespectively of context, or something else? Second, given a particular analysis, the goal is probably not to find all possible translations, even if that concept made sense, but what is it? To find the best translation? Even though it hardly makes sense to talk about all possible translations of a sentence, it might make sense to put the possible translations into groups, where the translations in the same group are more or less equivalent while there are essential differences between the groups. Would it in this case make sense to try to find (at least) one representative of all significantly different groups of translations? As with analysis, we could also here focus on context and only consider translations likely in the actual context. We should put these questions on our discussion agenda for 2005 and then focus in on what our goals shall be for 2006.

We have decided to enrich the parser with a stochastic ranking mechanism building on experiences from other LFG related projects. For this we need some training material in terms of an MRS or treebank. We plan a collaboration with the TrePil-project on this task. Then we will experiment with the ranking.

**WP 17.** Tree- and MRS bank for analysis ranking

**Responsible** Victoria

**WP 18.** Stochastic ranking of analyses

**Responsible** Helge

For the transfer and generation steps there will be more experiments with stochastic ranking carried out as part of Erik Velldal's Ph.D.-project carried out in collaboration with Stephan (supervisor).

In addition we shall include some hand craft. For example, in constructing the transfer rules there will be a constant balance between general

broad-coverage rules and more precise rules with a more restricted applicability.

Some pruning might also be done to the ERG and LKB-generator, and some of this might be done by hand, say to choose a consistent American (in contrast to a British) spelling. While we want all possible semantic representations corresponding to a string of words (sentence), we do not need all possible realizations of a particular semantic representation. In addition, Liv Ellingsen's project on soft constraints may add to the quality here.

Lexical ambiguity is a particular problem for machine translation, and we have so far put it into the background. There are several places in the translation process where this may be addressed. We may rely on the restricted domain and only propose one translation for each (open-class) word. Or we may try to do lexical disambiguation as part of the analysis (WSD), or we may rely on specifying contexts in the transfer rules directing the choice. Probably, we will use a combination of all three methods. In 2005 we will experiment with the different strategies and hopefully have a clearer strategy for 2006. Gunn Inger Lyses's project might contribute at this point. It is important that we early on get reasonably sized lexicas for the two languages and a transfer lexicon so we can experiment with the different strategies.

Ambiguity arises i.a. from homonymy in the lexicon, both complete homonymy between lemmas (all lemma forms being ambiguous), and homonymy between individual inflectional forms. We want to investigate the possibility of reducing the inefficiency resulting from this by devising a method for the automatic derivation of a lexicon version in which all homonyms that cannot be distinguished syntactically are collapsed into single entries with underspecified semantics. This will at least postpone inevitable bifurcations into alternative representations in the translation process, which ought to enhance overall efficiency. This task will be seen in connection with WP 14 (MRS packing).

**WP 19.** Automatic merging of homonyms in the parsing lexicon

**Responsible** Helge

## 4.7 Coverage

We cannot expect the demonstrator to give a full analysis and translation to more than, say, half of the strings in a test corpus. The XLE-system has the possibility of delivering a fragmented analysis where the full analysis fails. We have already started experiments with translations of these fragmented analysis. This work will continue as part of the work of all three steps: analysis, transfer, and generation. In particular, this will put demands on the

generator with respect to robustness and generation from far from perfect input.

Of course, we cannot expect the result from a fragmented syntactic analysis to be the correct or best one. Luckily, the fragmented analysis results in many alternative analyses, but then they have to be ranked. So far the ranking is not the best one. Moreover, the analysis can be very slow when it turns into fragmented mode and at several occasions it terminates without a result. A tagger might help. From a set of fragmented analyses those which correspond with the tagger in their word analysis should be preferred. To facilitate efficiency the best will be to tag the strings before they are sent to a fragmented analysis, such that tagger-assigned hypotheses are available to the grammar and XLE facilities for fragment selection. We should experiment with a set-up like the following. We run two copies of the XLE system with the Norwegian LFG, NorGram, in parallel. Both are working on the output from the same morphological analyzer. But while the first analysis runs directly on the output from the analyzer, the input to the other analysis is tagged first. This yields a direct possibility for comparing the two strategies. Maybe the result is that we will only use the tagged input, but we conjecture the following scenario. When the output from the untagged input is a full analysis, it will be preferred. Otherwise, we get a fragmented analysis from the tagged string which we use.

**WP 20.** Fragmented analysis with OB-tagged input

**Responsible** Paul and Helge

## 4.8 Lexical data bases

There are several different computational lexicas for Norwegian derived from more or less the same sources. It is a goal for LOGON to collect these resources, clean them up and build a generic tool for Norwegian language technology. LOGON will serve as a test bed for this resource. So far material has been collected and cleaned up, in particular we have got a better classification of lexemes and homonyms and, by using the concept of orthoclumps, lumped together lexemes which share meanings.

One goal now will be to develop good interfaces for future use in language technology. In particular, we will develop a good interface to the morphological analyzer used in the project so that updates to the lexicon can easily be reflected in the analyzer. Another goal is to put the lexicon with its semantic classification to use in the transfer process. Finally, aspects of packaging, documentation, and distribution will again be necessary parts of making sure that the lexical database is a re-usable, widely used resource for Norwegian language technology.

**WP 21.** Lexical data base

**Responsible** Jan Tore and Daniel

## 4.9 Evaluation

### WP 22. Evaluation

**Responsible** Torbjørn

## 5 Publications and other dissemination

Everybody who is employed by the project as a researcher is expected to publish on their project work. Principal investigators, even though they are not paid by the project, should, of course, also contribute to publications. A person responsible for a work package is also responsible for the publication of the results of that work package.

The most immediate channel for publication will be conference proceedings. But we expect also some of the more theoretical work, like WP 11 and WP 12, to result in journal publications. Of course, also some of the other work might yield results suitable for a journal publication.

In addition to this, we will produce a book documenting the whole project. This will be a collection of chapters with separate authors, where each chapter documents parts of the overall project.

### WP 23. Book

**Responsible** Jan Tore

The project will also arrange the conference of the European Association of Machine translation (EAMT) in 2006.

## 6 Progress plan

Much of the progress plan follows directly from the division into work packages above. We will not here make more specific time schedules for all the different work packages, but leave to the responsible persons of each work package to specify expected progress and present plans and results to the project work space. Some tasks are vital for other tasks and have to be completed by a certain date. We will mention some of the immediate ones below. Others may come up during the project period.

Much of the coordinated work will be organized around the demonstrator. We plan new releases of the demonstrator three times each year (April 1, September 1, and December 15) both in 2005 and 2006. The release of September 1, 2006, will be the main release, as we expect to present it more widely that fall already.

We will also consider the release of Sept. 1, 2005, the main release of 2005. A goal for this release will be to have a first version of the full lexicon of the training text. We do not expect a high quality on the full lexicon, but

we think that a first shot on a larger lexicon is necessary to get a better grasp on how to address several problems in the last year, in particular lexical ambiguity.

This means that some tasks will have the highest priority at the start of 2005. First of all we should get the training corpus as soon as possible and get it translated by April 1, 2005. Second, we shall as soon as possible extract the vocabulary and try to align it. The two grammars shall be extended to include as much as possible of the vocabulary. We should start with the simpler words, like the nouns. More complicated grammatical constructions and (non-)aligned translations will be considered during the whole project period.