Corpus based language technology for computer-assisted learning of Nordic languages: Squirrel Progress Report September 2001

Lars Borin, Stockholm
Lauri Carlson, Helsinki
Diana Santos, Oslo

1 Background and objectives of the project

1.1 Background

The most prevalent forms of information technology in computer-assisted language learning (CALL) applications are email, multimedia and computerised multiple-choice test forms. Until recently, the contribution of language technology (LT) to CALL has been next to nonexistent, although there is now a growing amount of work in the LT community on using natural language processing (NLP) technologies and speech technology for creating 'intelligent' CALL (ICALL) applications, where language analysis and learner modelling are used in order to analyse and respond properly to free-form linguistic input from the learners (e.g. answers to exercise questions) or even to score essays.

Text corpora and corpus tools, another kind of LT, are also used in CALL, indirectly in the form of so-called learner corpora, i.e. collections of second or foreign language texts produced by language learners. These corpora can be examined for interlanguage traits which in turn can inform language teaching.

There is also a more direct use of corpora in so-called "data-driven learning", where students use concordancers and search tools for empirical investigations of the use of words and constructions in large text materials. These tools are characterisable as language technology only in a very wide sense, since their linguistic sophistication is of a very low order (usually not even POS tagging is involved, let alone more sophisticated linguistic analysis).

It is important to note that the NLP technologies involved and large corpora are available on a very uneven basis. For English and possibly a few other languages there are both advanced language analysis tools and a number of multi-million word corpora. None of the Nordic languages are anywhere near as well equipped in this respect. Multilingual (parallel or comparable) corpora suitable for language learning are yet harder to come by, first because there are less of them and second because problems with securing copyrights and corpus preparation are multiplied.

There is a need for CALL applications utilising LT (LT-CALL) for Nordic languages, but before we can begin work on developing such applications, we must take stock of needs and resources, and take steps to secure easy access to them.

The aims of this feasibility study are to identify (i) the needs of a primary target group of language learners, (ii) the kinds of LT which we see as mature enough in the Nordic research community to lend themselves to fairly immediate use in CALL applications for a number of Nordic languages, and, finally, (iii) what LT resources and research would be needed in a slightly longer perspective for supporting computer-assisted learning of Nordic languages.

1.2 Target group

As a realistic primary target group for Nordic LT-CALL applications we have singled out (language teachers of) exchange students to Nordic universities and other institutions of higher education. This is an important and growing category of second-language learners in the Nordic countries, as a result of a general globalisation trend in higher education. To this category one might add students of Nordic languages as a foreign language, i.e. not in the country where the language is spoken, but e.g. at a German or Japanese university. For our purposes it is crucial that this category can be considered fairly homogenous as to their educational background and their computer literacy. This in turn means that we can concentrate more on the LT aspects of the study, avoiding having to simultaneously deal with a much more heterogenous learner group, such as immigrants.

Some LT-CALL resources suited for the exchange student category will undoubtedly turn out to be directly usable for some immigrant categories, as well as for minority children and even in native language instruction in schools. This is of course desirable, beacause in the longer term, all kinds of learners of Nordic languages should be considered, and not only exchange students.

1.3 Scope of the study

We see the acquisition and preparation of suitable text materials for language learners as a naturally delimited and realistic task where LT can be applied. A

constant problem in second and foreign language teaching for adults is a dearth of adequate training text material and up-to-date vocabularies, and the amount of teacher work that goes into continually preparing new materials. Here, LT can help out, using tools from corpus linguistics, information retrieval (IR), information extraction (IE), text mining, and possibly text summarisation. There are monolingual and multilingual corpus projects being pursued at many places in the Nordic countries, from simple corpus collection endeavours to cutting-edge research on tagging, parsing, alignment, lexicon extraction and machine learning from large corpora. There is also relevant work being pursued in the other areas mentioned.

This feasibility study is thus focusing primarily on how to process monolingual and multilingual text sources to produce adequate and up-to-date learning material for learners of Nordic languages as a second language. More specifically, we are investigating the issues of collecting and accurately classifying texts in Nordic languages as to their

1. language,

2. subject area, and

3. linguistic difficulty level.

For (1), very reliable methods exist already. For (2), there is relevant work on document classification and document segmentation in the IR community, which we can draw upon, although there is still work to do in this area. As to (3), much less has been done in LT or IR, while there is quite a lot of theoretical and practical work on this problem in applied linguistics, language pedagogy and didactics, and possibly also in human-computer interaction and cognitive psychology. For our purposes, (3) consists of two related problems:

3.1 By which criteria do we classify a text as “difficult” or “easy” for a language learner of a particular kind (vocabulary, syntactic complexity, semantic complexity, subject area, etc.)?

3.2 How do we operationalise these criteria so that they can be used for automatic text classification?

An obvious pilot application of this kind of document classification is a special purpose web search engine, i.e. language teachers and learners would get a tool which could collect and prepare desired amounts of training texts of the proper kind (domain/genre/dialect) and difficulty level (vocabulary/style/register) on demand off the WWW (see further section 3.1).

2 Preparations and stocktaking

The project started in fact only in the second quarter of 2001, due to the circumstance that the funding was awarded immediately before Christmas 2000, by which time the three project members had already filled their agendas for the first three months of this year. This means that the present report in effect documents the first half-year of work in the project.

At the start of the project Borin had been appointed to a temporary post as Section Head of Computational Linguistics in the Department of Linguistics at Stockholm University, with the practical consequence that his share of the project work is carried out at Stockholm instead of Uppsala.

The official project kickoff meeting took place on 27 March, 2001 in Oslo, hosted by Santos. For reasons of euphony, it was decided at that meeting not to use a project acronym; instead, we refer to the project by the 'nickname' Squirrel.[1]

2.1 Website

Santos has prepared a project website with public and internal webpages at <http://www.informatics.sintef.no/projects/CbLTCallNordicLang/squirrel.html>, as well as a project e-mail list for electronic discussions among the project members (<squirrel-team@informatics.sintef.no>). On the Squirrel website you will find brief presentations of the researchers and the project. On the internal pages (open to the project members and to observers; see section 2.2.2), work in progress is shown.

2.2 Taking stock of needs and resources

One stated aim of the project has been to assess the needs and resources in the area of computer-assisted learning of Nordic languages. Naturally, part of this information has come through a literature search (see section 2.2.1). Part will come through an e-mail list survey, which is yet to be conducted.

However, the most important part in our view – the part yielding the highest-grade information – is interaction with relevant people: second language acquisition researchers, language teachers, NLP researchers, and other stakeholders in this area.

[1] On the project webpages (see section 2.1) we say: "Squirrel, the project's working nickname, originates from our liking this small animal, an industrious gatherer of resources, of which we would like to provide a corresponding computational agent." This is a fair, even if brief, summary of the project aims.

2.2.1 Literature

A preliminary list of relevant literature – on intelligent computer-asssisted language learning, text difficulty, automatic summarising, and relevant work in the information retrieval and information extraction fields – was put together by Borin immediately at the start of the project, and later the other two project members have added entries and comments to this list. At present, it contains about 80 entries, but is still growing. The list is being prepared for publication on the Squirrel website, in two forms:

1. as a complete, uncommented reference list;

2. as a number of reviews of relevant literature in a particular field or from the point of view of a particular problem. Thus

    - Santos is preparing a review with the tentative title "Language technology for learning and teaching language – a preliminary overview", and she is co-supervising a Phd dissertation about computational linguistics methods of improving information retrieval on the Web (see section 4.4), with a corresponding review
    - Borin will review the literature on standards in the fields of e-learning and LT (see section 3.3), and he is supervising two Master's projects in Computational Linguistics, where literature on text classification and on text difficulty will be reviewed (see sections 3.1 and 3.2).

2.2.2 Project 'observers'

Already at the kickoff meeting on 27 March 2001, a number of Norwegian 'observers' were invited, representing parties that we had good reasons to believe would be interested in various aspects of computer-assisted learning of Nordic languages (particularly Norwegian in their case, of course). So far, the following researchers (in alphabetical order) have agreed to being observers of Squirrel:

- Jan Engh
- Asbjørn Følstad
- Hilde Hasselgård
- Stig Johansson
- Christian-Emil Ore

In addition to have a first insight about the linguistic resources and projects that could be relevant for a follow up of the Squirrel project, we received interesting feedback on our initial ideas, which can be summarized thus:

Linguistic resources We were warned that it is nowadays an established result that not too much weight should be given to frequency lists. In addition to frequency dictionaries, linguistic resources lacking for Nordic languages is what is called "Availability lexica", namely thesaurus-like organized lists of words that refer to areas of activity and/or subjects. Other relevant linguistic resources for teaching purposes that should not be subestimated were contrastive studies among the closely-related Scandinavian languages.

Museology Even though not directly related to language teaching, tools associated to museum collections – in as much as they presented culturally relevant realities – could be an exciting source of teaching material, also given their multimedia capabilities. An example is the kind of tools being developed by the Museum Project (see <http://www.muspro.uio.no/engelskomM.shtml>)

Usability Requirements elicitation should be done unbiasedly. Ideally, users (language teachers) should be involved in the project definition.

2.2.3 Other contacts

In Sweden, Borin has established contacts with second language acquisition researchers Björn Hammarberg at Stockholm University and Åke Viberg at Uppsala University. Further, contacts have been taken with teachers of Swedish as a second language, both at the university level (Berit Söderman in the Department of Scandinavian Languages, Uppsala University) and at the secondary school level (Hillevi Torell, Head Teacher of Swedish as a Second Language at Celsiusskolan, Uppsala). Borin will meet with Söderman and Viberg in the beginning of October to get information on which variables are likely to be important and thus needed to take into account in locating or adapting text material for learners of various backgrounds. At this meeting, we will also start a cooperation on collecting a second language Swedish learner corpus (see section 4.2).

In Danmark, Santos has long been collaborating with the VISL Project (South Denmark University at Odense) for Portuguese corpus-based research, and has gathered information about VISL's text-based teaching material as well as how they tackle issues of difficulty level.

In Norway, formal contact has been done with the Department of Linguistics of the University of Oslo (Ingeborg Kongslien), leading to further interviews early October.

2.2.4 Interview preparation

A common questionnaire or way to approach teachers of "Nordic language for foreigners" was also agreed early on, in order to provide a common basis for comparing the target community of the Squirrel project across the Nordic countries. A minimum of coherent information was identified, and formulated as a set of questions:

- What level do you teach (what are the classifications employed)?
- What sort of classes you give (how many students, how much oral vs. written participation, what kind of teaching material)?
- How many text material do you use (only textbook, specific texts, or tailored to the students needs and interests)?
- In case you use texts other than a standard textbook, how do you select the texts?
- Do you use electronic texts? If yes, how?
- Do you reuse the texts? Do you edit the texts?
- Do you get examples from the Web? Do you use Web repositories?

It is clear, though, that before trying to get these questions answered, some time should be allotted to hear the teachers explain how they usually work, for later scenario building. A possibly useful exercise has also been suggested: in common with the teachers, create some scenarios of how to use corpus-based LT in their work. This brainstorming should, however, be exercised with care, lest one ends up imposing our ideas on the users.

Then, the interviews should also provide us with relevant real life examples around which we could build our demo and start investigation.

Finally, it was agreed that inquiries about the teachers' opinions about difficulty levels, as well as their knowledge and actual use of these concepts/tools, should also be made during the interviews, given that measuring "text difficulty" was identified as one of the potentially problematic concepts to be looked into by the Squirrel project.

3 Project modules

3.1 Text classification and automated text search

Automated text search on the Internet, based on a relevant *text classification* is one of the main thrusts of the Squirrel project. Already at the kickoff meeting, a

fair amount of effort went into creating mockup user interfaces for such a "Text selector" application, where the relevant classification parameters were laid out in different ways. In the end, we came to the preliminary conclusion that we should aim for a fairly simple interface, at least for the first prototype, with stepwise access to more advanced options.

The mockup interface is now being filled with content. We pursue this work from several angles:

Borin is supervising a Master's project funded by Squirrel, where Kristina Nilsson, a Language Engineering student at Uppsala University, is building a prototype web crawler for locating texts according to specific requirements as to language (one of the official Nordic languages), subject area, and difficulty. At the moment, she is considering a number of freely available web search robots, as well as the possibility of using existing web search services, e.g. Google, building Squirrel-specific filtering on top of such a service.

Santos is also looking at the possibility of integrating a Web crawler (Ixkwic) and a language identifier (papagaio), developed at SINTEF by Paulo Rocha in the context of the Computational Processing of Portuguese project, in order to get texts on the same subject in the several Nordic languages.

3.2 Text adaptation

Borin is supervising a Master's project, where Anna Decker, a Computational Linguistics student at Stockholm University, is investigating the issue of text difficulty for second language learners – contrasting it with the more commonly encountered notion of text difficulty for native readers – and that of text simplification for language learners. The aim of the project is to lay the foundation for a language learner text simplification application. The idea of automatic simplification or adaptation of text comes from text summarisation, an IR technique for reducing textual information to its bare bones. Present-day text summarisation does not rely much upon linguistic knowledge to achieve its aim, and consequently the result is often a text that is usable for the purpose of getting the desired information from it, but not a good text on its own right. For language learners, the summary text ought to be grammatically acceptable, the evaluation of which requires fuller linguistic analysis (tagging, parsing). Decker will not be able to build a text simplification application (this is not even her assignment), but we believe that she may come up with some guidelines for developing such an application.

3.3 Standards in e-learning and LT

One issue which has emerged in the course of the project work is that of *standards*. Standards are created so that resources – both data and tools for working with the

data – can be reused. Language technology and other resources often represent very large investments in human effort and money, and standardisation is seen as a way of ensuring the long-term cost-effectiveness of these investments.

In the field of e-learning, standard formats are defined for all aspects of so-called 'instructional management systems'. Thus, not only educational content formats are agreed upon, but also course structure formats, test formats, as well as how their interaction with recordkeeping systems used in education should take place. There is a number of organisations working on standards in the e-learning area, the most important ones being IMS (Instructional Management System Inc; <http://www.imsproject.org/>), IEEE's LTSC (Learning Technology Standards Committee; <http://ltsc.ieee.org/>), the American Department of Defence ADL (Advanced Distributed Learning; <http://www.adlnet.org/>) initiative, and the European ARIADNE (Alliance of Remote Instructional Authoring and Distribution Networks for Europe) project. Standards being developed by these and other bodies include educational metadata (LOM), test formats (IMS QTI), content packaging formats (IMS content packaging), and modular courseware (ADL SCORM). At least some of these standards are rapidly gaining acceptance in the e-learning industry. Thus, learning applications – such as the ones developed or proposed by the Squirrel project – will need to support them in order to be viable in the long term.

There is also ongoing standardisation work in the language technology community, e.g. in organisations such as EAGLES (Expert Advisory Group on Language Engineering Standards), CES (Corpus Encoding Standard), etc.

The field targeted by Squirrel, Intelligent CALL – i.e., computer-assisted language learning incorporating and utilising language technology (in a non-trivial way) – does not really seem to be part of either of those two worlds, however. This means that little consideration is taken of ICALL in the ongoing standardisation efforts in the field of e-learning or in the field of language technology.

However, standardisation is an issue which needs to be addressed if we want the resources that we create to be reusable in whole or in (modular) parts. Borin presented a paper on this issue at the *13th Nordic Conference on Computational Linguistics* in Uppsala in May 2001 ("Making ends meet: which e-learning standards for Intelligent CALL?"). This presentation is being written up as a Squirrel report to be published on the project website.

3.4 The other Nordic languages

The primary target languages for the project are, naturally enough, the Nordic languages represented by the three project members, i.e. Finnish, Norwegian (both Bokmål and Nynorsk), and Swedish. In a slightly longer perspective, we also intend to include the other official languages of the Nordic countries. Eventually,

however, we should consider working with all the languages having some kind of official status in the Nordic area, thus including both languages covered under the *European charter for regional or minority languages*, as ratified by the respective Nordic countries – e.g. Saami, Yiddish, or Romani – and the languages of the extended Nordic area, i.e. the Baltic states and the westernmost part of Russia.

As we said in the project proposal (repeated above in section 1.1), language resources of various kinds are very unevenly distributed among the world's languages. The kind of web text search facility being prototyped in Squirrel depends – among other things – on there being web-based materials available in sufficient amounts, and it is an empirical question whether this holds for all the languages in question.

Borin is currently making a preliminary study of this issue for Finnish Romani, i.e. he tries to make an assessment of the amount and kind of publicly available text material on the Internet in this language. The results of this study will be presented at the *8th Nordic Conference on Bilingualism*, to be held in Rinkeby, Sweden, in November 2001 ("Babe in the woods or new kid on the block? Minority languages and ICT: the case of Finnish Romani"). This presentation will be written up as a Squirrel report to be published on the project website.

4 Collaborations and overlaps with other related work

There is a good deal of ongoing related research work where the Squirrel investigators are either directly involved as investigators ('overlaps') or indirectly involved through collaboration ('collaborations'). Below, we mention some of the most relevant such overlaps and collaborations.

4.1 Methodology for multilingual corpus studies

Santos has been doing a contrastive study of of the prepositions *com* (Portuguese) and *med* (Norwegian), based on a bilingual dictionary, annotated corpora of the two languages, and parallel corpora including the two languages (both the OMC, including translations of the same English texts into the two languages, and ordinary parallel corpora having as "opposite language" English). This study is methodologically interesting, as Santos concentrates on exploring what can and cannot be done with such resources as already exist.

Borin, together with Klas Prütz (Department of Linguistics, Uppsala University), has worked on a method for investigating syntactic transfer phenomena in learner language. The study was done on learner English, so the material is not

directly relevant for Squirrel, but the methodology is. This consists in comparing three corpora (native L1, learner L2, and native L2) with respect to part-of-speech tag sequences. This is ongoing work, but an account of how the same method was used for investigating *translationese* will be published in the proceedings of the *Computational Linguistics in the Netherlands 2000* conference ("Through a glass darkly: part-of-speech distribution in original and translated text" by Lars Borin and Klas Prütz).

4.2 Swedish learner corpus creation

Borin has initiated the collection of text material for the creation of a (written) Swedish learner corpus, in collaboration with Berit Söderman (Department of Scandinavian Languages, Uppsala University) and Åke Viberg (Department of Linguistics, Uppsala University). Essays written by SL Swedish students at the university will be collected by teachers, together with relevant information about the learners. Initially, this will happen only at Uppsala University, but contacts have been taken with representatives for the Royal Institute of Technology (KTH) in Stockholm, where similar SL Swedish courses are held, and with a representative for the secondary school system in Uppsala (Hillevi Torell), both of which will be followed up during the fall.

Learner corpora of Nordic languages (Swedish in this case) are important resources for the kinds of applications being developed in the Squirrel project. The availability of such corpora will allow us – among other things – to relate text difficulty not only to features in the target text, but also to features in the learners' own target language production, which in turn can be more finely correlated (at least sometimes) to such variables as the learner's native language, educational background, etc.

4.3 Second language Swedish grammar checking

Borin has started a collaboration with a research group at KTH in Stockholm, who are interested in developing their existing grammar checker for (native) Swedish into a grammar checker for non-native writers of Swedish. The collaboration is at a preliminary stage, but it has resulted in a project proposal for Vinnova, as well as given the impetus to Borin for initiating the creation of a Swedish learner corpus (see section 4.2). Further, one of the researchers in the group, Ola Knutsson, and Borin are co-supervising a Master's project on grammar checking for learner Swedish by Anna Staerner, a Language Engineering student at Uppsala University.

4.4 Web information retrieval

Santos is supervising a PhD student, Rachel Aires, in NLP-based information retrieval in the Web. One of the early results of this is an overview of NLP methods for improving both the task of finding out the subject of a text and the purpose of a query. The two areas that are more relevant for Squirrel are 1) text categorisation methods and 2) usability of query systems as a whole.

4.5 Parallel corpora with corrected versions

Some work of a very preliminary nature has been pursued by Santos concerning the use of (second language learner) corpora aligned with their corrections – and also with their translations –, with the purpose of identifying and measuring the possible advantages of engaging in a larger project to develop this kind of system. Languages involved were Portuguese written by Norwegians (with Norwegian as original text) and English written by Brazilians (with Portuguese as original text).

5 Looking ahead

According to the plan, the prototype text collector (see section 3.1) will be fully specified before the end of the year. This will allow us to turn to the next item on our agenda, viz. to study how current work in LT on parallel and comparable corpora and corpus tools (such as sentence and word alignment tools for parallel corpora, and bilingual lexicon extraction tools for parallel and comparable corpora) can be applied to the problem of producing up-to-date glossaries for selected subject areas.

Via the ongoing work on text difficulty (see section 3.2), we also hope to be able to at least begin to take on the idea of text adaptation.

By the end of the project we will have a much clearer picture of where we stand and what still needs to be done, i.e. we will document our findings in the form of reports, as well as endeavour to formulate new projects in the area of LT-CALL.