*LREC 2002. Third International Conference on Language Resources and Evaluation. Proceedings. Vol II*. Las Palmas, Spain: ELRA. 2002. 411–418.

# Living off the land:
# The Web as a source of practice texts for learners of less prevalent languages

**Kristina Nilsson and Lars Borin**

Computational Linguistics, Department of Linguistics,
Stockholm University, SE-106 91 Stockholm, Sweden
and
Department of Linguistics, Uppsala University,
Box 527, SE-751 20 Uppsala, Sweden

krinil@stp.ling.uu.se, lars.borin@ling.su.se, lars.borin@ling.uu.se

## Abstract

This study focuses on how to automatically locate text sources published on the World Wide Web in order to produce adequate and up-to-date learning materials for second language learners of Nordic languages. The Web is an excellent source of authentic text materials. However, the large amount of information available on the Web makes search services necessary. Hence, we are developing *Squirrel*, a prototype Web meta-search service, described in this paper, which collects text material in the Nordic languages according to language, topic and difficulty level. Our primary target group consists of exchange students to Nordic institutions of higher education, and their language teachers, although in the longer perspective, we would also like to be able to do something for minority language communities. We describe the basic implementation of Squirrel, and present preliminary results from trying it out. Finally we discuss the (lack of) Web resources in less prevalent languages, and how we imagine that applications like Squirrel could fit into a second or foreign language learning situation.

## 1. Introduction

The Squirrel project – or officially: “Corpus based language technology for computer-assisted learning of Nordic languages” – is a feasibility study funded by the Nordic Council of Ministers.

The study focuses on how to automatically locate text sources published on the World Wide Web in order to produce adequate and up-to-date learning materials for second language learners of Nordic languages. Research on reading in a second language indicates that the use of authentic text material for this purpose is both more adequate and more motivating than using specially prepared courseware. Of course, the location and selection of authentic text material is more demanding on the teacher than the use of ready-made courseware. Hence, one aim of the Squirrel project is to propose how techniques from the fields of Computational Linguistics, Information Retrieval (see, e.g., Baeza-Yates and Ribiero-Neto (1999)), and Information Refinement (Olsson, 2002) can be brought to bear on this task in order to make it easier for second-language teachers in particular, but also for their students, to locate and select suitable authentic text material.

The Web is an excellent source of authentic text materials: “One of the most interesting aspects of computerized text is that almost all of it is authentic discourse.” (Cobb and Stevens, 1996, 118). However, the large amount of information available on the Web makes search services necessary. Hence, we are developing a prototype Web meta-search service, described in this paper, which collects text material in the Nordic languages according to topic and difficulty level. Our work relies on findings from a number of research areas in the fields mentioned above, including Web information retrieval, automatic language identification, and text categorization.

Our primary target group consists of exchange students to Nordic institutions of higher education, and their language teachers. Students of Nordic languages as a foreign language at institutions outside of the Nordic region might also be added. This target group can be considered to be fairly homogeneous as to their educational background and computer literacy. However, in the longer perspective we would also like to be able to do something for minority language communities, such as the Sami and Finnish Roma communities discussed in the next section and in section 5.2., below.

## 2. The Nordic languages

The Nordic languages are:

(1) official state languages Danish, Finnish, Icelandic, Norwegian-Bokmål, Norwegian-Nynorsk, and Swedish;

(2) official regional languages Faroese and Greenlandic Inuit;

(3) officially recognized minority languages Meänkieli (Torne Valley Finnish), Romani, Sami,[1] and Yiddish (for each of which at least one Nordic country has signed the *European charter for regional or minority languages* (Council of Europe, 1992)).

[1] Actually, both Romani and Sami consist of a number of (more or less mutually intelligible) languages or dialects (depending on your criteria).

Although all of these are to be counted as small languages on a world scale, in terms of their number of speakers, at least category (1) languages are represented with millions of documents on the World Wide Web.[2]

For example, a search for the highly frequent Icelandic function word *að* (‘to’) (following Ghani et al. (2001)) with the Google search engine returns about 1,310,000 Web pages. The other category (1) languages yield similar results, as shown in Table 1.

| Language | Results |
|---|---|
| Icelandic *að* (‘to’) | 1,310,000 |
| Norwegian-Nynorsk *ikkje* (‘not’) | 230,000 |
| Norwegian-Bokmål *ikke* (‘not’) | 1,620,000 |
| Danish *ikke* (‘not’) | 1,470,000 |
| Finnish *että* (‘that’) | 1,660,000 |
| Swedish *är* (‘is’) | 8,020,000 |

Table 1: Google search results

Thus, there are good grounds for believing that the Web could serve as a source of text material for learners of these languages. Consequently, we need a search service which will locate texts according to their language, topic, and difficulty level.

# 3. Toward a text search solution

Information retrieval on the Web is different from classic information retrieval in that queries must be answered from the index without access to the text. Most Web search engines wage an uphill battle against the rapidly increasing amount of Web pages and frequent updating, which conspire to make gathering and maintaining data using the prevalent crawler-indexer architecture almost impossible. (Baeza-Yates and Ribiero-Neto, 1999)

The development of a search engine from scratch lies well beyond the scope of this project. Instead, the Web search service Evreka[3] is used in a meta-search approach, where topic query terms are extracted from example documents provided by the user.

The Squirrel application has been implemented in Perl 5. Modules in the libwww-library have been used to handle Web requests and retrieval, and parsing of HTML documents for links, contents and meta data such as summaries and keywords.[4]

## 3.1. Running Squirrel

Squirrel, the application described in this article, was designed to explore the Web as a source of material for language learners. Initial results are encouraging, even though the application remains to be formally tested and evaluated.

In this prototype version of Squirrel, each search is initialized by the user supplying an example text, in the form of an URL. The HTML document this URL refers to is collected, and the text is evaluated as to language, word count, and readability score. This information is presented to the user, as well as ten possible query terms extracted from the example text. The user choses which of the terms should be sent to the search engine Evreka. The documents referred to by the 60 highest ranking URLs returned from Evreka are collected and evaluated one at a time. The language of each document is compared to the language of the example text, and if they are the same, a readability score is computed.

When all the documents have been evaluated, the user is presented with a list of possible URLs. For each URL, the number of words in the document, the readability score, the document title, and a summary are also presented.

## 3.2. Extracting topic query terms

In order to extract possible query terms from the example document, the probabilistic term frequency of each word is computed. Firstly, the text is tokenized by converting the text from a string of characters into a list of words. All characters are changed to lower case, and punctuation marks and digits are removed. The words in the remaining text string are identified as items separated by spaces. Finally, words found in a list of stopwords are eliminated, in order to filter out high frequency words such as articles, prepositions, and conjunctions, that most likely have nothing to do with the topic of the text. (Baeza-Yates and Ribiero-Neto, 1999)

The HTML structure of the document is also checked for meta-keywords supplied by the author of the document. If there are no meta-keywords, the ten most frequent words are presented to the user as possible query terms. If there are meta-keywords in the document, a comparison is made with the probabilistic term frequency list. The words that are present in both lists are presented as the most likely query terms, and the most frequent words not in the meta-keyword list are used to fill up the top 10. The final selection of query terms is made by the user.

## 3.3. HTML document retrieval and parsing

The HTML documents are retrieved by a user agent, an interface layer between the Squirrel application and the Web , through which remote servers can be accessed. The user agent handles communication with the remote server by sending a request for the URL and collecting the content, the HTML document.

The retrieved HTML document is parsed for the text content, that is the HTML tags are removed, leaving only the text segments. This is done by breaking up the HTML documents into different segments much like a browser would, and storing the text segments. The document is also summarized by collecting elements of the HTML structure that can be used to describe the topic of the document. The meta description, the document title, headlines, and text segments written in bold face are examples of such elements.

The parsing of the HTML document is quite important, since the quality of the result has a great impact on further

[2] For category (3) languages, however, the situation is quite different; Sami and Finnish Romani are discussed in section 5.2. below.

[3] Evreka URL: `http://evreka.passagen.se/`

[4] The libwww-library and modules can be downloaded from the Comprehensive Perl Archive Network, CPAN. URL: `http://www.cpan.org/`

processing, such as the language identification and the difficulty level assessment of the text. HTML documents are difficult to parse since the specifications for HTML change continually, and HTML markup is often both incorrectly and inconsistently written. There are many special cases to consider; embedded JavaScript code, frame sets, style tags, and comments extending over multiple lines to name but a few.

As mentioned above, the search terms extracted from the example document are sent to the search engine Evreka. The HTML of the returned result page is parsed to find the links, which are divided in two groups: the direct links and the links to the remaining result pages (if any). The latter are then accessed one by one, and parsed for all direct links. These links are added to the original list of direct links. This list is then processed, one item at a time, by accessing each document with the user agent. Each document is evaluated by examining the HTML structure and the text content.

### 3.4. Written language identification

Choosing among a number of good methods available for written language identification (Mathusamy and Spitz, 1996), we selected the N-gram-based Text Categorization method by Cavnar and Trenkle (1994). This method compares rankings of the most frequent n-grams to assign a document to a predefined category. It is based on Zipf's Law, which states that the n-th most common word in a human language text occurs with a frequency inversely proportional to n. This also holds for the frequency of occurrence of n-grams, the implication being that documents within the same category should have similar n-gram frequency distributions.

This text categorization method has been implemented as TextCat (van Noord, 2001),[5] a written language identification program. TextCat currently supports 69 languages, including Icelandic, Norwegian,[6] Danish, Swedish and Finnish, and can be modified and redistributed under the terms of the GNU General Public Licence.

Since TextCat does not differentiate between the two variants of Norwegian, Nynorsk and Bokmål, new language models have been created for this application. These models were created from collections of texts for each language, consisting of roughly 25,000 words each. The texts were collected from online news sites, and used as input to the TextCat application for creating new language models.

A language filter based on TextCat was used by Ghani et al. (2001) in CorpusBuilder, a system which automatically generates Web search queries for collecting documents in minority languages. The performance of the filter was tested by native speaker evaluation, giving a precision of nearly 100 percent. Informal performance tests of the Squirrel application indicate similar encouraging results: the precision of TextCat language identification ranges from 94 percent and upwards (see Table 2).

| Language | Precision |
|---|---|
| Icelandic | 100% |
| Norwegian-Nynorsk | 100% |
| Norwegian-Bokmål | 96,8% |
| Danish | 94,1% |
| Swedish | 100% |
| Finnish | 100% |

Table 2: Precision of the written language identification by TextCat

### 3.5. Difficulty level ≈ readability

The term *readability* is used to describe text characteristics which can be used to predict how easy or difficult texts are to read and to understand. Discourse structure, complex phrase structures, and the amount of abstract and difficult terminology are examples of such text characteristics. (Alderson and Urquhart, 1984)

Formulas for measuring text readability are often based on more easily calculated surface linguistic features such as word length and sentence length, which are believed to correlate with these 'deeper' characteristics. The Flesch-Kincaid readability metric, which is the U.S. Department of Defense standard, was developed to test the readability of military training manuals. The metric measures the average number of words per sentence and the average number of syllables per word, with manually assigned parameters. (Si and Callan, 2000) Other methods focus on reader characteristics, such as the reading ability of each student represented by the grade average (Mikk, 1995; Mikk and Elts, 1999), and on concept difficulty, based on the hypothesis that this can be captured by statistical language models learned from actual corpora (Si and Callan, 2000).

Research by Strother and Ulijn (1987) show that lexical rewriting can increase reading comprehension in English as a second language used in science and technology education, but that no such benefits can be accomplished by simplifying the syntax. Strother and Ulijn suggest that the focus in second language reading education should be on vocabulary and on the development of reading skills. Ghadirian (2002) describes a computer-based method for incremental introduction to topic-specific vocabulary. A collection of authentic texts of a certain topic are sorted according to the percentage of high frequency words, thus ensuring that each new text contains a suitable amount of new terms and concept.

In determining the difficulty level of a text, one must keep in mind the situation in which the text is to be used, and the method of analysis should be chosen depending on which linguistic difficulties are regarded as important. For example, if the lexical complexity is the most important factor then frequency lists can be used. Within the Intelligent Web-based Interactive Language Learning Project (IWILL), a lexical difficulty filter has been designed to filter corpus search concordancing results. The filter uses a frequency list and a function which allows the user to set a threshold level for the filter. Concordance results are thus filtered according to the chosen threshold level and pre-

[5] TextCat Language Guesser URL: `http://odur.let.rug.nl/~vannord/TextCat/`

[6] Although it does not distinguish between Norwegian-Bokmål and Norwegian-Nynorsk; see below.

sented to the user, giving the user comprehensible authentic language data. (Wible et al., 2000)

The use of readability formulas has been criticized because they only measure surface linguistic features, which is regarded as primitive and in some cases even misleading. One of the reasons for this criticism is that readability formulas have been used prescriptively as a style checker tool for writers, even though they were originally designed for descriptive use. Studies by Karlgren (2000) and Platzack (1973) show that if the objective is to efficiently grade texts according to difficulty, readability formulas work well as long as they are used descriptively.

Platzack (1973) studied the effect of syntactic complexity on reading comprehension, and came to the conclusion that since only a few of the all the possible different syntactic structures occur frequently in any given text, one may as well use surface symptoms to grade texts according to readability. This conclusion is supported by a study by Karlgren (2000), where the most important factors in readability testing were found to be word length and sentence length. Among other factors tested in the study were the average number of nouns, adverbs, prepositions, and pronouns.

It is important, however, to differentiate between readability and suitability, since the readability score of a text only measures surface features. Suitability must be determined by the teacher and/or reader, and this is especially important if the text is collected from the Web. Not only can the content be inappropriate (for a number of reasons), but it can also be too difficult if the reader lacks the necessary background knowledge. (James, 1987)

One further caveat is that readability measures have typically been used with the native reader in mind, whereas their (at least direct) applicability to second and foreign language reading has not been systematically investigated, as far as we know. Regardless of this, we have adopted readability as an easily calculated possible first approximation of text difficulty level for second language learners. This choice must be subject to evaluation, of course (see section 5., below).

In the Squirrel application, the readability of the document is measured with the Lix readability formula due to Björnsson (1968), which is based on average sentence length and percentage of long words (more than 6 characters). This formula has been used for readability testing on a large number of children's books, novels, text books, and newspaper articles in 6 languages (Swedish, German, Danish, English, Finnish, and French). It has also been used in less comprehensive tests on newspaper articles in Norwegian-Bokmål, Norwegian-Nynorsk, Italian, Spanish, Portuguese, and Russian. (Björnsson and af Segerstad, 1979)

## 4. Initial results

Since the Squirrel application collects texts according to language, difficulty level, and topic, the overall performance of the system depends on the combined results of the language identification, the readability assessment, and the document relevance.

Traditionally, precision and recall have been used to measure the performance of information retrieval systems. It has been argued that since these concepts implicitly assume that the users want a complete set of relevant documents, they are not well suited for the Web. (Nielsen, 1999) In modified form, however, precision and recall can be applied to Web search. Web users are interested in short response times, and the precision as well as the recall of the URLs listed on the first page of retrieved documents, since search engine users seldom view more than the top 10 or 20 results of a search. (Kobayashi and Takeda, 2000)

The results of a search with the search engine Evreka are ranked according to how many times the query terms occur in each document, and where in the HTML structure the query terms are located (for example in the document title). Ranking is also influenced by the number of references from other Web sites. Evreka uses the FAST global index AlltheWeb.[7] However, the precision of the relevance of the documents depends not only on the performance of the search engine, but on the query terms as well. Therefore, the topic query term extraction method used in the Squirrel application must be evaluated.

As mentioned in section 3.4, the TextCat language identification performs well. Results of informal tests of the readability assessment and the relevance of the documents are also encouraging, but the prototype must be further tested and evaluated.

### 4.1. Example test results

The example document used in this test is a review in Swedish of the film "Sagan om ringen - Ringens brödraskap" (Lord of the Rings - The Fellowship of the Ring).

```
Styrka kan finnas i de minsta av saker.
En av 1900-talets populäraste litterära
verk, Tolkiens Sagan om ringen, blir nu
det nya århundradets största filmhän-
delse.  Ett hett efterlängtat epos och
actionäventyr som regisserats av Peter
Jackson.  Inspelningarna har gjorts i
Nya Zeeland's böljande landskap.  Sagan
om ringen är första delen i trilogin
Härskarringen.
Sagan om ringen berättar historien om
den unge hoben Frodo som ärver en till-
synes oskyldig ring.  Frodo upptäcker
att ringens ursprunglige skapare, den
onde trollkarlen Sauron, desperat letar
efter ringen.  För det är en ring med
mycket onda krafter som kan göra det
möjligt för Sauron att förslava invå-
narna i riket som kallas Midgård.
```

Figure 1: The first two paragraphs of the example document

There are 938 words in the text, and the readability score is 43. This information about the example document was presented to the user:

[7] AllTheWeb FAST Search URL: http://www.alltheweb.com

```
LANGUAGE: Swedish
NO OF WORDS: 938
READABILITY: 43
DOCUMENT SUMMARY: Bio.nu - Filminfo.
10 MOST FREQUENT TERMS (incl meta keywords):
1 ringen
2 sagan
3 peter
4 jackson
5 bio
6 filmen
7 the
8 frodo
9 dess
10 nya
```

From this list, the query terms “ringen”, “sagan”, and “frodo” were chosen. This query resulted in 60 URLs, which is the maximum number of URLs retrieved by Squirrel. 15 of these were unreachable, but the remaining 45 were successfully retrieved. Of these 45 documents, 38 were in Swedish.

The documents presented below are of the same language as the example document, and of the same number of words or more, sorted in increasing order based upon the readability score.

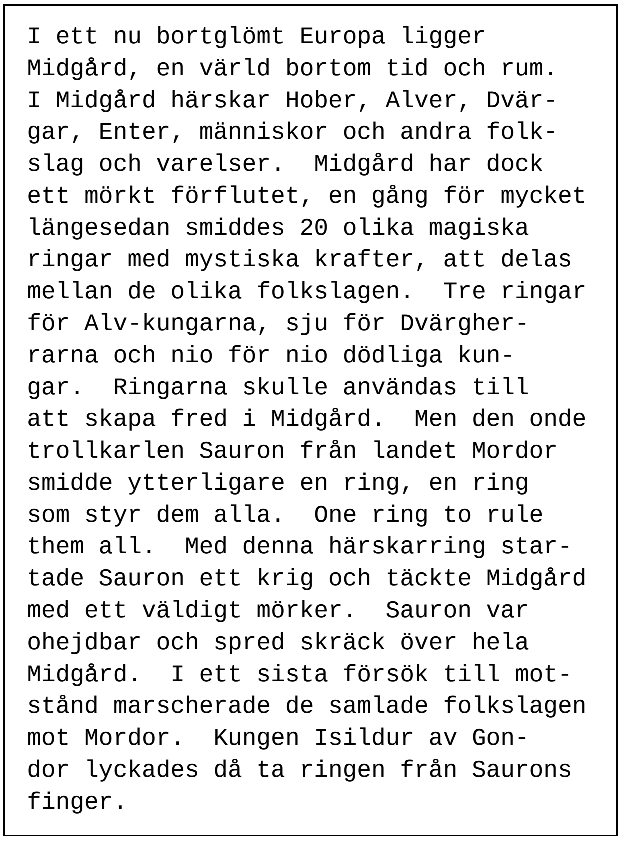
```
URL: http://www.moviemix.nu/filmrec.asp?ID=191
NO OF WORDS: 1784
READABILITY: 34
DOCUMENT TITLE: SAGAN OM RINGEN - Moviemix med filmrecensioner från video, bio och dvd
DOCUMENT SUMMARY: Moviemix har dom färskaste filmrecensionerna

URL: http://exolite-network.com/exolite/avdelningar/recensioner/sagan_om_ringen.asp
NO OF WORDS: 1066
READABILITY: 36
DOCUMENT TITLE: Exolite Network - Recensioner - Sagan om Ringen
DOCUMENT SUMMARY: En recension på den efterlängtade filmen om härskar ringen!

URL: http://www.amosmagasin.com/ArticlePages/200110/24/20011024151348_Adminstallet200/20011024151348_Adminstallet200.dbp.html
NO OF WORDS: 1442
READABILITY: 36
DOCUMENT TITLE: AMOS MAGASIN
DOCUMENT SUMMARY: Man skulle varit fluga

URL: http://iacobaeus.com/boklista/
NO OF WORDS: 961
READABILITY: 39
DOCUMENT TITLE: Den feta boklistan
DOCUMENT SUMMARY: Feta boklistan

URL: http://mediaarkivet.com/filmrec.asp?typ=1&id=357
NO OF WORDS: 1033
READABILITY: 40
DOCUMENT TITLE: mediaarkivet.com - filmrecensioner och musikrecensioner
DOCUMENT SUMMARY: MediaArkivet.com - Media åt folket!

URL: http://www.allamedia.com/spelfilm/artikel_158.shtml
NO OF WORDS: 944
READABILITY: 41
DOCUMENT TITLE: Sagan om Ringen
DOCUMENT SUMMARY: En av de mest uppskattade filmerna genom tiderna.

URL: http://vujer.com/recensioner.php?rid=531
NO OF WORDS: 1461
READABILITY: 43
DOCUMENT TITLE: Vujer.com | Sagan om ringen
DOCUMENT SUMMARY: Du är här:
```

All of these documents are film reviews, except the document titled “Den feta boklistan”, which is a book review. Below are the first two paragraphs from the document titled “SAGAN OM RINGEN - Moviemix med filmrecensioner från video, bio och dvd”, which had the lowest readability score:

> När jag var tio år fick jag höra talas om “Sagan Om Ringen”, bläddrade i den och tänkte: “Det här måste jag läsa någon gång.” Nästan tjugo år senare har “någon gång” fortfarande inte inträffat, men då har sagan hunnit bli en film istället. Vad har inte sagts redan om “Lord Of The Rings” som inte tål att upprepas? Ingen aning.
>
> Skulle ändå vilja be om att få inleda denna recension med att jag inte håller med om att filmen är gjord enbart bara för att tjäna pengar; visst kostade den enormt mycket att göra och filmbolaget New Line Cinema satsade allt de ägde, på en i och för sig rätt säker hand, men de vågade åtminstone spela högt.

Figure 2: The first two paragraphs of the text with the lowest readability score.

The document titled “Vujer.com | Sagan om ringen” had the highest readability score, which corresponds with the score of the example document.



I ett nu bortglömt Europa ligger Midgård, en värld bortom tid och rum. I Midgård härskar Hober, Alver, Dvärgar, Enter, människor och andra folkslag och varelser. Midgård har dock ett mörkt förflutet, en gång för mycket längesedan smiddes 20 olika magiska ringar med mystiska krafter, att delas mellan de olika folkslagen. Tre ringar för Alv-kungarna, sju för Dvärgherrarna och nio för nio dödliga kungar. Ringarna skulle användas till att skapa fred i Midgård. Men den onde trollkarlen Sauron från landet Mordor smidde ytterligare en ring, en ring som styr dem alla. One ring to rule them all. Med denna härskarring startade Sauron ett krig och täckte Midgård med ett väldigt mörker. Sauron var ohejdbar och spred skräck över hela Midgård. I ett sista försök till motstånd marscherade de samlade folkslagen mot Mordor. Kungen Isildur av Gondor lyckades då ta ringen från Saurons finger.

Figure 3: The first paragraph of the text with the highest readability score

# 5. Looking ahead

In this section, we will bring up for discussion a mixed bag of issues, which have arisen at various points during our work with the Squirrel prototype: (1) the issue of how to make the prototype into a useful tool for language learning, (2) the issue of the unequal status of the various Nordic languages and their availability on the Web, (3) the issue of future applications, being logical or more far-fetched continuations of our work in Squirrel, and, finally, (4) the issue of how we envision the use of such applications in a second or foreign language learning situation.

## 5.1. Improvements to the prototype

This prototype could be improved in many ways. Aside from a user interface, there are several functions that would increase the usability, for example if the user were allowed to give a number of example texts, or to give examples by “cut and paste” or by defining search paths to locally stored text files. The user should also be able to add new search terms other than the terms suggested by Squirrel, since the query term extraction method used, probabilistic term frequency, is far from perfect.

Tests and evaluation with language learners and teachers – planned for later in 2002 – will show whether this tool will be as useful as we think it could be in second language learning. This testing and evaluation will of course add items to this list of possible improvements.

## 5.2. ‘Less prevalent languages’ – a mixed lot

In contrast to the category (1) Nordic languages (according to the classification given in section 1.) – which belong to the so-called *high-density* languages – the category (3) languages present a very different picture. As part of the Squirrel project, Borin (2001) made a separate investigation of the use of Sami and Finnish Romani on the WWW. It turned out that not only do Sami and Finnish Romani show much less of a presence on the Web, compared to the category (1) languages, but also that they are very different in comparison to each other.

First some background information on the groups in question. The Sami number about 80,000 (50,000 in Norway, 20,000 in Sweden, 10,000 in Finland, and 2,000 in Russia). The Finnish Roma number about 15,000 (10–13,000 in Finland, and 3–4,000 in Sweden). Sami has a longer tradition of writing and literacy than Finnish Romani.[8] There are no reliable figures as to how many first-language speakers of Sami and Finnish Romani there are.

Lund (2000) tries to assess the use of Sami in Web pages of public institutions, organizations, etc., in Norway, Sweden and Finland, having some kind of connection to Sami issues. He finds that barely 25% of the information content in the web sites investigated by him (29 sites out of the 66 he investigated had some information in Sami, but the amount and quality was extremely varying), is available in one or another of the Sami languages.

Borin made a number of searches with Google, using Finnish Romani function words *hin* ‘is, are’ and *dola* ‘this, the’, as well as different inflectional forms of the ethnonym *Kaale* ‘Finnish Rom(a)’. He came up with altogether 2 web pages containing texts in Finnish Romani (out of about 10 pages altogether, most containing only a few words of Finnish Romani, e.g. in the form of cited phrases).

There are probably a number of reasons for this state of affairs. Lund (2000) blames technology. According to him, the cards are doubly stacked against Sami in the world of computers: Firstly, there is an English-language dominance, from which even major European languages suffer.´ Secondly, the most widespread character encoding in the Nordic area, Latin-1, does not cover the whole Sami alphabet. In other words: It is difficult to use Sami on computers.

In the case of Finnish Romani, matters are a bit more complex. There is a standard orthography (Ortografiakomitea, 1971), but its use is not that common in print. Here, we have look for less tangible cultural clues to the low presence of this language on the Web. In brief, the following are probable contributing factors:

(1) Finnish Roma as a group have a comparatively low educational level, with concomitant low proficiency in writing and reading, as well as low exposure to computers;

(2) They live in a traditionally oral culture, while the Web – despite the multimedia content – is a mostly written language medium (see also Gough and Bock (2001));

[8] E.g., there are chairs in Sami and Sami departments in universities in Norway, Sweden, and Finland, but no higher academic institutions where Finnish Romani is taught.

(3) There is a feeling that the language belongs solely to the community of its traditional speakers, the Finnish Roma, and that information about it should not be available to outsiders, i.e. it is a 'secret language', while the WWW is an extremely open medium.[9]

Now, neither Sami nor Finnish Romani will be likely target languages for our chosen group of learners, i.e. exchange students in the Nordic countries. On the other hand, the need for authentic text material is not confined to second language learning. Sami or Roma children receiving native language instruction in school could also benefit from the kind of tool that we are trying to develop in the Squirrel project. Needless to say, the applicability and usefulness of the tool are not restricted to the Nordic area or the Nordic languages.

### 5.3. Ideas for the future

The work described here could be extended in various ways:

(1) Refining the Lix formula with vocabulary (mainly corpus frequency; see Wible et al. (2000)), and syntax data (from tagging and partial parsing);

(2) Extracting word lists and other information from the texts, possibly in combination with bilingual dictionaries, or (more speculatively) comparable texts in the native language of the learner (cf. Rapp (1995); Fung and Yee (1998); Diab and Finch (2000));

(3) Adapting texts automatically, i.e. simplifying difficult texts, in a process akin to automatic summarization;

(4) Using the learners' own written output (i.e., learner corpus data) for finding texts on a suitable difficulty level – somewhere (just) above that evidenced in the learner's production;

(5) Automatically preparing exercises and tests from texts located and downloaded from the Web. Cobb and Stevens (1996) propose a number of such exercises and tests, at least part of which could be generated automatically using computational linguistics techniques and language technology tools.

### 5.4. Language learning with Squirrel and its ilk

An application like the Squirrel text search engine, as well as the other, more ambitious conceivable applications presented in the previous section, should all be seen mainly as *tools for the language teacher*. There is no realistic scenario where language learners, especially beginners, could use such tools without the guidance of an experienced teacher of the language in question. We can take some of the drudgery out of a task such as wading through large amounts of material in order to locate a few suitable texts, but the ultimate judgement as to which of those texts are suitable and which should be discarded must be left in the hands of a human, and the language learners – almost by definition – do not have the requisite knowledge and experience to make such a judgement (cf. Borin and Gustavsson (2000)).

Consequently, we consider it very unlikely that we will ever see the 'science fiction future' when computer-assisted language learning applications – even ones equipped with 'intelligence' in the form of language technology – will be anything more than a complement to the human teacher.

## 6. Acknowledgements

The work reported herein was carried out as part of the project *Corpus based language technology for computer-assisted learning of Nordic languages*, in the framework of the Nordic Language Technology Research Program 2000–2004 (Holmboe, 2001), funded by the Nordic Council of Ministers through Nordisk Forskeruddannelsesakademi (NorFA).

## 7. References

J. C. Alderson and A. H. Urquhart, editors. 1984. *Reading in a Foreign Language*. Harlow: Longman.

Ricardo Baeza-Yates and Bethier Ribiero-Neto. 1999. *Modern Information Retrieval*. ACM Press Books, Addison-Wesley.

C. H. Björnsson and Birgit Hård af Segerstad. 1979. *Lix på franska och tio andra språk*. Stockholm: Pegagogiskt centrum, Stockholms skolförvaltning.

C. H. Björnsson. 1968. *Läsbarhet*. Lund: Liber.

Lars Borin and Sara Gustavsson. 2000. Separating the chaff from the wheat: Creating evaluation standards for web-based language training resources. In Khaldoun Zreik, editor, *Learning's W.W.W. Web Based Learning, Wireless Based Learning, Web Mining. Proceedings of CAPS'3*, pages 127–138, Paris. Europia.

Lars Borin. 2001. Babe in the woods or new kid on the block? Minority languages and ICT: the case of Finnish Romani. Presentation at the 8th Nordic Conference on Bilingualism, November 1-3, 2001, Stockholm - Rinkeby.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94), Las Vegas, Nevada, U.S.A.*, pages 161–175, World Wide Web, `http://citeseer.nj.nec.com/68861.html`, April. Downloaded 2001-07-08.

[9] This means that there could, in principle, be a great amount of Finnish Romani material on the Web, but hidden from the casual user and from search engines. This is unlikely, however.

Relevant in this context is also that there has been some – but altogether too little – debate in the literature about linguistic and other work on some community made by outsiders to that community – e.g. fieldworking linguists, anthropologists, sociologists, or missionaries – and to what extent and under which circumstances such work constitutes 'appropriation', and even 'violation of the rights' of the community in question (Howes, 1996; Östman, 2000). These issues become more, not less, relevant when we discuss the creation of language technology resources for small languages (often minority languages threatened by extinction) on the basis of resources already existing for major languages.

Tom Cobb and Vance Stevens. 1996. A principled consideration of computers and reading in a second language. In Martha C. Pennington, editor, *The Power of CALL*, pages 115–136. Athelstan, Houston, TX.

Council of Europe. 1992. European charter for regional or minority languages. European Treaties ETS No. 148. World Wide Web, `http://www.coe.fr/eng/legaltext/148e.htm`, Downloaded 1999-11-29.

Mona Diab and Steve Finch. 2000. A statistical word-level translation model for comparable corpora. Technical Report LAMP-TR-048/UMIACS-TR-2000-41, Department of Computer Science, University of Maryland.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of COLING-ACL'98*, pages 414–420, Montréal. ACL, Université de Montréal.

Sina Ghadirian. 2002. Providing controlled exposure to target vocabulary through the screening and arranging of texts. *Language Learning & Technology*, 6(1):147–164.

Rayid Ghani, Rosie Jones, and Dunja Mladenic. 2001. Building minority language corpora by learning to generate web search queries. Technical Report CMU-CALD-01-100, Carnegie Mellon University Center for Automated Learning and Discovery. World Wide Web, `http://citeseer.nj.nec.com/444684.html`, Downloaded 2001-09-04.

David H. Gough and Zannie Bock. 2001. Alternative perspectives on orality, literacy and education: a view from South Africa. *Journal of Multilingual and Multicultural Development*, 22(2):95–111.

Henrik Holmboe, editor. 2001. *Nordisk sprogteknologi. Nordic Language Technology*. Museum Tusculanums Forlag, Københavns Universitet, Copenhagen.

David Howes. 1996. Cultural appropriation and resistance in the American Southwest: decommodifying 'indianness'. In David Howes, editor, *Cross-Cultural Consumption. Global Markets, Local Realities*, pages 138–160. Routledge, London.

Mark O. James. 1987. ESL reading pedagogy: Implications of schema-theoretical research. In Joanna Devine, Patricia L. Carrell, and David E. Eskey, editors, *Research in Reading in English as a Second Language*. Washington D.C., U.S.A.

Jussi Karlgren. 2000. *Stylistic Experiments for Information Retrieval*. Ph.D. thesis, Department of Linguistics, Stockholm University.

M. Kobayashi and K. Takeda. 2000. Information Retrieval on the Web. Technical Report RT0347, IBM.

Svein Lund. 2000. Ingen samisk på internett? World Wide Web, `http://hjem.sol.no/sl1015/sami/inetsan1.htm`. Downloaded 2001-10-30.

Yeshwant K. Mathusamy and Lawrence Spitz. 1996. Automatic language identification. In Ronald A. Cole et al., editor, *Survey of the State of the Art in Human Language Technology*. World Wide Web, `http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html`. Downloaded 2001-06-11.

Jaan Mikk and Jaanus Elts. 1999. A reading comprehension formula of reader and text characteristics. *Journal of Quantitative Linguistics*, 6(3):214–221.

Jaan Mikk. 1995. Methods for determining optimal readability of texts. *Journal of Quantitative Linguistics*, 2(2):125–132.

Jakob Nielsen. 1999. User interface directions for the web. *Communications of the ACM*, 42(1).

Fredrik Olsson. 2002. Requirements and design considerations for an open and general architecture for information refinement. *RUUL, Reports from Uppsala University, Department of Linguistics*, 35.

Ortografiakomitea. 1971. Mustalaiskielen ortografiakomitean mietintö. Komiteamietintö 1971: A 27. Helsinki.

Jan-Ola Östman. 2000. Ethics and appropriation – with special reference to Hwalbáy. In Frances Karttunen and Jan-Ola Östman, editors, *Issues of Minority Peoples*, pages 37–60. Department of General Linguistics, University of Helsinki. Publications No. 31.

Christer Platzack. 1973. *Språket och läsbarheten*. Lund: CWK Gleerup Bokförlag.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting of the ACL*. ACL.

Luo Si and Jamie Callan. 2000. A statistical model for scientific readability. World Wide Web, `http://citeseer.nj.nec.com/454751.html`. Downloaded 2001-09-18.

Judith B. Strother and Jan M. Ulijn. 1987. Does syntactic rewriting affect english for science and technology (E S T) text comprehension? In Joanna Devine, Patricia L. Carrell, and David E. Eskey, editors, *Research in Reading in English as a Second Language*. Washington D.C., U.S.A.

Gertjan van Noord. 2001. Textcat language guesser. World Wide Web, `http://odur.let.rug.nl/~vannoord/TextCat/`. Downloaded 2001-09-04.

David Wible, Chin-Hwa Kuo, Feng yi Chien, and Chih-Chiang Wang. 2000. Adjusting corpus searches for learners' level: Filtering results for frequency. Presentation at TALC2000, Graz, Austria, July.