

University of Gothenburg
Language Technology Programme
May, 2008

**FROM CORPUS TO LANGUAGE CLASSROOM:
reusing Stockholm Umeå Corpus
in a vocabulary exercise generator
SCORVEX**

Master Thesis, 30 points
Author: Elena Volodina

Supervisor: Lars Borin

May, 2008

Abstract

In this master thesis the focus has been made on the evaluation of Stockholm Umeå Corpus (SUC) as a source of teaching materials for learners of Swedish as a Second language. The evaluation has been carried out both practically and theoretically. On the theoretical side, readability tests have been run on all SUC texts to analyze whether appropriate texts can be automatically selected for each proficiency level. To make readability analysis more “vocabulary aware” lexical frequency profile of each text has been collected, analyzed and embedded into the final readability score assigned to each text. SUC has proven to be a rich source of texts of different proficiency levels appropriate for language training purposes. Advantages and disadvantages of SUC as a source of pedagogical materials have been identified in the course of work.

On the practical side, as a side effect of the theoretical analysis, a pedagogical tool SCORVEX (Swedish CORpus-based Vocabulary EXercise generator) has been designed and implemented. The existing modules of SCORVEX demonstrate to which extent it is possible to generate pedagogically acceptable vocabulary items with SUC as the only language resource. I am demonstrating in the thesis how wordbank items, multiple choice items and c-tests can be automatically generated for a specified proficiency level, word frequency band and a specified wordclass. In yes/no items potential words are generated on the basis of existing morphemes. All the four modules are therefore “language-aware”. Accessing frequency data obtained from SUC is the pre-requisite for the exercise generation, whereas SUC text archive is the only source of texts, sentences and words for vocabulary items.

This thesis can hopefully wake interest among teachers to test this generator in real-life conditions and maybe even convince some teachers in the usefulness of this pedagogical tool. The numerous ways for further development of this software are outlined in the paper.

CONTENTS

ABSTRACT	1
CONTENTS	2
List of Tables	4
List of Figures	4
List of Abbreviations	5
1. INTRODUCTION	6
1.1 VOCABULARY ACQUISITION – A FEW WORDS	6
1.2 EXERCISE GENERATORS - BACKGROUND AND RELATED RESEARCH	7
1.3 IDEA AND CENTRAL ISSUES OF THIS ESSAY	8
1.4 METHOD	11
1.5 STRUCTURE OF THE THESIS	11
1.6 NOVELTY AND APPLICABILITY	12
2. ICALL FOR SWEDISH: OVERVIEW.....	13
2.1 CALL - OVERVIEW OF DEVELOPMENT	13
2.2 ICALL - OVERVIEW OF DEVELOPMENT	15
2.3 SWEDISH AS A SECOND/FOREIGN LANGUAGE.....	19
2.3.1 Teaching/Testing Swedish as a Second Language.....	19
2.3.2 Research within Swedish as a Second Language. Linguistic & Pedagogical Perspectives	20
CALL APPLICATIONS FOR SWEDISH AS L2	22
2.5 ICALL APPLICATIONS FOR SWEDISH AS L2	22
2.5.1 GRIM	23
2.5.2 IT-based Collaborative Learning in Grammar (ITG).....	24
2.5.3 VISL - Visual Interactive Syntax Learning.....	25
2.5.4 Ville & DEAL.....	26
2.5.5 ARTUR.....	26
2.5.6 VocabTool	27
2.5.7 Lingus	28
2.5.8 Wordfinder.....	28
2.5.9 Squirrel.....	29
2.5.10 Didax.....	29
2.5.11 Other projects.....	29
2.6 NL RESOURCES AND NLP TOOLS FOR SWEDISH.....	30
3. USE OF CORPUS IN THE EXERCISE GENERATOR.....	31
3.1 GENERAL ON CORPORA IN SECOND LANGUAGE ACQUISITION	31
3.2 OVERVIEW OF SWEDISH CORPORA	32
3.3 GENERAL ON SUC AND ITS ROLE IN THE EXERCISE GENERATOR.....	33
3.4 SOME WORDS ON THE NOTIONS OF “WORD” AND “LEMMA”	36
3.5 SUC AS A SOURCE OF FREQUENCY INFORMATION.....	37
3.5.1 The FL in yes/no items.....	40
3.5.2 The FL in automatic selection of target vocabulary items from texts	40
3.5.3 The FL in selection of distractors for multiple-choice items	43
3.5.4 The FL in search of authentic texts. LFP calculation.....	43
3.6 SUC AS A SOURCE OF AUTHENTIC EXAMPLES	45
3.6.1 Readability Indices.....	46
3.6.2 Lexical Difficulty Measures.....	47
3.6.3 Test setting.....	49
3.6.4 Test results, generalizations and conclusions.....	52
3.6.5 Algorithm for text selection.	57
3.6.6 Algorithm for sentence selection	58

4. VOCABULARY GENERATOR – PEDAGOGICAL PREREQUISITES, THEORETICAL QUESTIONS AND DESIGN.....	60
4.1 GENERAL INFORMATION ON THE GAP CLOZE TEST ITEMS.....	60
4.2 COMPUTER-ASSISTED GENERATION OF C-TESTS.....	62
4.2.1 Automatic selection of target words.....	62
4.2.2 Automatic text and sentence selection.....	62
4.2.3 Correction for grammar and spelling.....	63
4.2.4 Calculation of the score.....	64
4.2.4 Examples of automatically generated c-items.....	66
4.3 COMPUTER-ASSISTED GENERATION OF MULTIPLE-CHOICE ITEMS.....	68
4.3.1 Selection of Distractors.....	69
4.3.2 Selection of Sentences/Texts.....	71
4.3.3 Scoring Procedures.....	71
4.3.4 Examples of automatically generated multiple-choice items.....	71
4.4 COMPUTER-ASSISTED GENERATION OF WORD BANK ITEMS.....	73
4.4.1 General Information on Word Bank Items.....	73
4.4.2 Examples of automatically generated word bank items.....	74
4.5 SWEDISH VOCABULARY SIZE TEST.....	78
4.5.1 General Information on the Test Design.....	78
4.5.2 Generation of Potential Swedish Words.....	79
4.5.3 Calculation of the score.....	80
5. CONCLUDING REMARKS.....	84
5.1 SUC – ADVANTAGES AND DISADVANTAGES.....	84
5.2 FUTURE OF THE SYSTEM.....	87
5.2.1 Towards the specificity of existing exercises.....	87
5.2.2 Towards expanding of the system.....	88
5.2.3 Towards a better user interface.....	88
5.2.4 Towards improved presentation and user adaptability.....	89
5.2.5 Experiments and tests.....	89
5.2.6 Other areas of application of the generator.....	90
5.3 RESULTS.....	90
REFERENCES.....	93
APPENDICES.....	101
APPENDIX 1. CORPORA OF SWEDISH.....	101
Corpora of Written Swedish (non-commercial).....	101
Corpora of Spoken Swedish.....	102
Learner Corpora.....	103
APPENDIX 2. FUNCTION WORDS IN 8 FREQUENCY BANDS.....	105
APPENDIX 3. DIAGRAMS OF FB DISTRIBUTION PER EACH LIX LEVEL, AVERAGE VALUES.....	111
APPENDIX 4. TEXTS USED FOR READABILITY GRADING BY HUMAN READERS.....	117
APPENDIX 5. SWEDISH CONSONANT CLUSTERS.....	126
APPENDIX 6. IMPLEMENTATION OF SCORVEX MODULES – SOME FACTS.....	127
Implementation of C-test Items.....	127
Implementation of Multiple-Choice Items.....	129
Implementation of Word Bank Items.....	131
Implementation of Swedish Vocabulary Size Test.....	132

List of Tables

Table 1. Overview over ICALL applications for Swedish as L2	23
Table 2. Structure of the base vocabulary pool.	38
Table 3. List of POS tags used in base vocabulary pool	39
Table 4. List of POS tags used for manual markup of word lists	43
Table 5. List of functional wordclasses.....	50
Table 6. Number of functional words per frequency band.....	51
Table 7. Average values per LIX level and frequency band	52
Table 8. Value span for each FB and LIX level.....	53
Table 9. Easiest and most difficult texts ordered by LIX.....	54
Table 10. Easiest and most difficult texts ordered by LFP-score	54
Table 11. Easiest and most difficult texts ordered by LexLIX.....	54
Table 12. Ranking of texts graded for difficulty by human readers from easiest to difficult	56
Table 13. LIX, LexLIX and LFP-scores in the 9 human-graded texts	56
Table 14. Examples of automatically selected distractors.	70
Table 15. Aspects of word knowledge.	91
Table 16. Standard deviation of FB, LD and LV values per each level.	116

List of Figures

Figure 1. Excerpt from SUC. An example of an annotated sentence	35
Figure 2. Schematic representation of corpus use in the exercise generator	36
Figure 3. Manual selection of texts with manual mark-up	40
Figure 4. Manual selection of texts with automatic mark-up	41
Figure 5. Automatic selection of texts with an automatic mark-up.....	42
Figure 6. Creating exercises from a list of words.....	42
Figure 7. Creating exercises from an automatically selected list of target words.....	42
Figure 8. Automatically collecting a list of distractors for multiple-choice items.....	43
Figure 10. SUC-sentence index. Content of the file “folkskola .NCU .txt”	58
Figure 11. C-test Module, user interface of the authoring tool.	65
Figure 12. Multiple Choice Module, user interface of the authoring tool	69
Figure 13. User Interface of the Word Bank Items Module	74
Figure 14. Stimulus-response matrix taken from (Huitbregtse et al. 2002).....	81
Figure 15. Swedish Total Vocabulary Test – User interface.....	83
Figure 16. UML-scheme for the C-test Module.....	127
Figure 17. UML-scheme for the Multiple Choice Module.....	130
Figure 18. UML-scheme for Word Bank Items Module	132
Figure 19. UML scheme for the module on Swedish Vocabulary Size Test.....	133

List of Examples of automatically generated items

Example 1. C-test: automatically selected nouns for training in a text of intermediate level.....	66
Example 2. C-test: Automatically selected words from FB 3000-4000 in a text of pre-intermediate level.	66
Example 3. Multiple-choice items: automatic search for adverbs in a text of pre-intermediate level	71
Example 4. Multiple-choice items: automatically selected nouns for training in sentences of intermediate level.	72
Example 5. Word bank items: exercise created on the base of a list of manually typed words (5 times the same word). Variant 1.....	74
Example 6. Word bank items: exercise created on the base of a list of manually typed words (5 times the same word). Variant 2.....	75
Example 7. Word bank items: differentiating between different forms of pronouns. Target vocabulary has been typed in by the user (not automatically generated!)	75
Example 8. Word bank items: differentiating between different forms of participles. Target vocabulary has been typed in by the user (not automatically generated!)	75
Example 9. Word bank items: automatically selected words from FB2.....	76
Example 10. Word bank items: automatically selected text for level 3 with automatically marked words from FB3	76
Example 11. Word bank items: automatically selected prepositions for training in sentences	76
Example 12. Word bank items: automatically selected prepositions in an automatically selected text.....	77

List of Abbreviations

ASU	Andraspråkets StrukturUtveckling (Second Language Structural Development), learner corpus
CALL	Computer-Assisted Language Learning
CEF	Common European Framework of references for language (language proficiency levels)
CES	Corpus Encoding Standard
CL	Computational Linguistics
DCG	Definite Clause Grammar
DIALANG	Diagnostic Of Languages, CALL Software
DM	Dialogue Manager
FB	Frequency Band
FDG	Functional Dependency Grammar
FL	Frequency List
GSLC	Gothenburg Spoken Language Corpus
GU	University of Gothenburg
ICALL	Intelligent Computer-Assisted Language Learning
IMS	Instructional Management Systems
ITG	IT-based Collaborative Learning in Grammar (ICALL software)
ITS	Intelligent Tutoring System
KTH	Kungliga Tekniska Högskolan (Royal Technical University)
L1	Native Language (Mother Tongue)
L2	Second Language
LD	Lexical Density
LE	Language Engineering
LexLIX	readability index LIX corrected for lexical complexity
LFP	Lexical Frequency Profile
LIX	readability index for Swedish: Läsbarhets Index
LR	Language Resource
LT	Language Technologies
LV	Lexical Variation
NL	Natural Language
NLP	Natural Language Processing
PoS / POS	Part-of-Speech
QTI	Question and Test Interoperability, IMS
SFI	Swedish For Immigrants
SLA	Second Language Acquisition
SUC	Stockholm Umeå Corpus
SVANTE	Svenska AndraspråkTExter, a corpus of learner texts in Swedish
SWEDEX	SWEDish EXamination
TEI	Text Encoding Initiative
TISUS	Test in Swedish for University Students
VISL	Visual Interactive Syntax Learning (ICALL software)

1. Introduction

Natural Language Processing (NLP) technologies are effectively used in many areas of human life, including the area of intelligent Computer-Assisted Language Learning (CALL). The latter focuses ordinarily on learners and their needs, rather than teachers and their needs. With existing language resources like tagged corpora, wordnets, lexicons, part-of-speech taggers, syntactic parsers etc. it is a shame that language teachers still have to produce a lot of learning materials and tests manually.

1.1 Vocabulary acquisition – a few words

Words are recognized as essential building blocks of the language. Language users that know the grammar of a language cannot explain themselves if they do not know words. However, knowing words without knowledge of grammar can help communicate ideas. Lexical competence is therefore important for language acquisition and effective communication.

Native speakers develop their lexical competence in early childhood, filling the existing blanks in response to new experiences as the need arises, i.e. incidentally. For second language learners the picture is more complicated: vocabulary acquisition is a conscious and time-consuming process that has to be supported by specially designed activities for more effective progress. Vocabulary can be acquired in different ways – through conscious learning (e.g. memorizing lists of words, doing vocabulary exercises, using target vocabulary in speech or writing) or through incidental learning (e.g. reading, listening). The fact remains though: vocabulary acquisition should be assisted if the learner is to develop good lexical competence in a fast and effective way (Nation & Waring 1997; Read 2000; Ma & Kelly 2006).

It is a fact supported by many researchers in second language acquisition that testing and assessing lexical knowledge falls into two traditional dimensions: breadth and depth (Gyllstad 2004; Zareva 2005). There are even other frameworks for vocabulary assessment, consisting of three and even four dimensions (Read 2000; Zareva 2005).

Breadth, otherwise called discrete-point approach, evaluates the receptive knowledge of words based on recall and recognition and deals with assessing the size of a learner's vocabulary. Words are used out of context¹ with supportive clues. Multiple-choice exercises, definition exercises and other types of exercises with supportive choices belong to this group.

Depth, otherwise called assessing quality of vocabulary knowledge, evaluates whether the learner knows all shades of meaning of a word and its typical contexts. This type of assessment is characterized by a communicative approach, i.e. vocabulary is not viewed as a separate construct, but rather as a natural part of language as a whole. This ability to use words productively in speech and writing is sometimes even referred to as receptive-productive knowledge of a word (Read 2000; Zareva 2005).

¹ The question is how to define context: sentence-long, text-long or even longer.

The second approach (depth or receptive-productive one) is gaining more popularity since it is argued that words acquire their meanings in context and should therefore be assessed and trained in context. However, though the limitations of discrete-point assessment have been recognized for a long time, multiple-choice tests, definition exercises and gapped sentences continue to be the most popular and the most widely-used formats of vocabulary assessment (Read 2000; Gyllstad 2004). There are several factors that are of importance: such tests are easy to administer, they are objective in nature and there is a long tradition with well-established procedures in how to produce and assess such tests. More important is, though, that such exercises do not exclude indirect/incidental learning of words so characteristic of native speakers. On the opposite, exercises of breadth type support incidental learning providing at the same time more training and rendering effectiveness to learning vocabulary.

1.2 Exercise generators - background and related research

The area of automated question generation presents a number of interesting research questions and is a focus of some current research (that deals however mostly with English as a source language).

There is a variety of approaches to this problem. A number of researchers studying the automated question generation use conceptual structures, others use ontological engineering, Directed Acyclic Graph (DAG) knowledge structures based on semantic networks (Li & Sambasivam 2005) and others. Here I will exemplify three approaches.

Jonathan C. Brown et al. (2005) make use of WordNet to generate six types of vocabulary assessment exercises: definition, synonym, antonym, hyperonym, hyponym, and cloze questions. They start from a prepared wordlist of relevant vocabulary items, thus pre-identifying which words to use in automatically generated exercises. As for the semantic annotation of polysemantic or homonymous words, they either do that manually or go for the most frequent items according to the WordNet frequency statistics.

Exercises are presented either with wordbanks or in the form of multiple-choice questions. Their approach in collecting distractors is based on selecting words of the same wordclass and similar frequency (Brown, Frishkoff & Eskenazi 2005).

Ruslan Mitkov et al. (2003) describe a computer-aided procedure for generating multiple-choice tests from electronic instructional documents. The main NLP techniques used in their system are term extraction, shallow parsing, a set of transformation rules and word sense disambiguation alongside with the use of such language resources as corpus and WordNet. The system works in several steps:

The first step is term extraction, which consists in identifying key concepts that serve as “anchors” for questions. This is done by identifying noun phrases with help of the FDG shallow parser. Next, the frequency of noun phrases in a domain-specific corpus is compared and those terms that are domain-specific (i.e. having frequency over a certain threshold) are selected as key terms.

Selection of distractors is the second step. It is done by consulting WordNet and retrieving synsets/hyperonyms for the “anchor”-word. The coordinated terms and

hyponyms to the anchor's hyperonym are selected as distractors. The preference is given to those distractors that appear in the domain corpus.

Question generation is the third step, which consists in applying transformation rules to the statements containing an anchor. A question is generated with minimal change of the original wording. The system consults agreement rules to ensure grammaticality of generated questions (Mitkov & Ha 2003).

Hideobu Kunichika et al. (Kunichika, Minoru, Tsukasa & Akira 2003; Kunichika, Minoru, Tsukasa & Akira 2005) describe a system aimed at Japanese learners of English where questions and answers are generated on the basis of a learner text. The system contains even a function for giving hints to a student if his/her previous answer is incorrect. The questions are generated on the basis of syntactic and semantic information extracted from the text, and as many questions as possible are generated using transformation rules. The system generates four types of questions:

- (a) a general question generated from one sentence;
- (b) a special question generated from one sentence;
- (c) a general question generated from more than one sentence;
- (d) a special question generated from more than one sentence;

Syntactic and semantic information from the stories is extracted using a method based on Definite Clause Grammar (DCG). Syntactic information is presented in a syntactic tree, containing information on parts of speech, modification relations, feature structure, etc. Semantic information shows time and space relations so that the information on time order of events can be easily retrieved and relations expressed by pronouns can be referenced to content words or relevant context (Kunichika et al. 2003; Kunichika et al. 2005).

It is worth mentioning that there exist a number of commercially available programs generating vocabulary exercises. To name a few, Exercise Generator developed by Oxford University Press (<http://www.clarity.com.hk/program/exercisegenerator.htm>), MCQ developed by Intcom (<http://www.intcom.se/MCQ/Overview.htm>), Exercise Generator Multi-Language produced by World of Reading, Ltd. (<http://www.wor.com/shopping/shopexd.asp?id=4193>). The common trait of them all is that they are language-independent, i.e. they take a text in any language (or almost any language) and with the help of some algorithms transform it into a number of exercises, like gapfill, jumbled words, sentence matching, misspelled words, etc. No text analysis or other NLP technologies are used to create exercises². These programs view texts as a bag of words and work for several European languages, including Swedish (MCQ).

1.3 Idea and central issues of this essay

Knowing a word implies knowledge of different aspects of the word and its usage. Nation (2001) identifies the following aspects, all of them having receptive and productive knowledge (modified and grouped by the author):

Form: spoken (recognition in speech, pronunciation)

² Information comes from personal communication over telephone or e-mail and personal testing of the demo versions

- written (recognition in texts, spelling)
- word parts (morphology: inflection, derivation, word-building)
- Meaning: form and meaning
- concept and referents
- associations
- Use: grammatical functions
- collocations
- constraints on use: register/frequency/etc.

As has already been mentioned, there exist a number of systems that can generate vocabulary exercises – mostly for English. Very few of them are based on NLP technologies and language resources. The general tendency is to use pre-programmed exercises or manipulate texts without text analysis (e.g. lemmatization, etc.). It is however obvious that a language learning tool that can be adjusted to the learner level and need can help teachers individualize language teaching and save teachers' precious time on creation of exercises.

In this Master Thesis I am trying to study both theoretically and practically possibilities that Stockholm Umeå Corpus offer for computer-assisted generation of vocabulary training exercises. The main purpose of this Master Thesis has originally been set out to answer the three principal theoretical questions:

- What aspects of word knowledge (see the list above) can be trained by computer-generated exercises based on SUC? To what effect?
- What aspects cannot be automatically generated from SUC and why? Which other NLP resources/tools are needed to cover the rest of word knowledge aspects? Are those tools/resources available?
- What resources are unavailable today to make automated generation of such exercises possible?

The practical evaluation of SUC has been carried out through implementation of an exercise and test generator³. The authoring tool (or exercise generator) has been given the name SCORVEX which stands for Swedish CORpus-based Vocabulary EXercise generator. The original ambition was to create a complete comprehensible system for vocabulary training. With time the ambition had to be readjusted to the time limits. The implemented part consists of:

- total vocabulary size measure of the type Paul Meara produces manually (see <http://www.swan.ac.uk/cals/calsres/lognostics.htm>, choose X_Lex: The Swansea Vocabulary Levels Test
- exercise generator part including multiple-choice exercises, wordbank items and cloze exercises.

³ In this thesis the system that has been designed and implemented by the author of this thesis is called interchangeably as: SCORVEX, the exercise generator, the (implemented) generator, the authoring tool, the system, the program, etc. – to avoid tautology.

On the way a number of interesting problems (i.e. the ones that could not be solved through the use of the existing NLP technologies and resources) have been studied, but not necessarily solved:

- Automatic identification of relevant words for training in learner texts versus manual marking of such words;
- Automatic selection of texts of an appropriate proficiency level;
- Automatic selection of sentences with target words of an appropriate proficiency level;

What distinguishes SCORVEX from the majority of commercial exercise generators is the use of NL resources that makes it possible to

(a) use base vocabulary pool for adjusted frequency information (Forsbom 2006). This information is necessary for selecting wordlists according to different learner levels, for selecting distractors for multiple-choice items, for total vocabulary size measure test, etc.;

(b) analyze a text and automatically identify relevant target items for the learner level in the learner texts;

(c) create a list of basic word forms or even lemmas of target words in a text supplying their wordclasses. This information is used as the basis for generation of all the exercises and tests;

(d) select a text of appropriate learner difficulty for creation of an exercise;

(e) select a number of authentic sentences with target vocabulary for wordbank items and cloze exercises from SUC;

The programme is also able to work independently of a text.

Generated exercises can be saved for regular paper use, i.e. in text format so far. In the future, one more format is planned to be implemented – QTI format – a standard proposed for creation of tests and exercises – for online use and automatic correction.

A number of interesting questions has been left for future work. Among those the following can be named:

- exercises based on morphological information since there is no available NLP resource with words organized in word-families or tagged for word-building morphological constituents;
- exercises on collocations for the same reason as above (no reliable NLP technology for identifying collocations in a text);
- feedback on learner performance. This question needs deeper research than I have had time for during this master thesis;
- analysis of short answers in the form of free writing for reliable correction of the answers as requiring deeper research;
- frequency lists based on spoken language (based on GSLC) and their lemmatization alternatively deriving base forms of the words;

- hyperlinking (of relevant target) words in the text to the entries in a dictionary collecting even concordance information and “best examples” for each lexical item. Hyperlinking in itself would probably not present a lot of problems. The selection of suitable concordance examples, however, is a complex question requiring deeper research.

1.4 Method

The starting point has always been the exercise type and its pedagogical prerequisites. Available technologies and resources have been analyzed to see which ones can help generation of the desired vocabulary item best. Interesting or difficult computational and linguistic problems were identified as the work progressed; some of them solved in the process of work and have been described in this paper.

The implemented generator functions as a practical test of the theoretical analysis. Algorithms for each exercise type have been described in a section specifically devoted to each particular exercise type.

1.5 Structure of the thesis

This thesis consists of six chapters.

The first chapter is an overview of the most important aspects of vocabulary in second language acquisition, some background research in the relevant area, and a few words on the main ideas of the thesis.

The second chapter is devoted to the overview over ICALL area – intelligent computer-assisted language learning in general and for Swedish in particular.

The third chapter deals with the questions around use of Stockholm Umeå Corpus in SCORVEX, in particular how frequency information is used in the automated generation of exercises, and how authentic texts and sentences are selected according to the user proficiency level.

Chapter four is a description of the particular exercise types and the linguistic and computational issues connected with them. Screenshots of the authoring tool are provided here as well as in appendix 6 where the implemented system, its design and most important algorithms are described. Some examples of the automatically generated exercises are provided.

Chapter five summarizes advantages and disadvantages of SUC as a source of vocabulary training exercises. I also summarize the results of the study, draw conclusions, describe some possible future development of the system and comment on what other resources are needed to cover the aspects of vocabulary learning that have not been covered by this generator.

A number of appendices are provided as a support to the information described in the chapters.

1.6 Novelty and applicability

Automatic generation of exercises is no novelty in itself. There are, however, no existing generators of vocabulary items for Swedish known to me, that take language aspects like word frequencies, wordclasses etc. into consideration, use NL resources and that can automatically provide learner texts of appropriate level.

The generator is supposed to be included as a part of the system ITG (Språkdata) and should therefore be open for use for those who have access to ITG.

The types of vocabulary tests and exercises that can be generated by SCORVEX can be used either for progress tests, for continuous training of target vocabulary or for assessment (diagnostic and final). In its nature this generator can produce:

- (a) general frequency-band based tests. The main use of those is for pre-tests, placement into level groups and evaluation of total vocabulary size of the learner;
- (b) syllabus-based exercises since the vocabulary scope can be predefined by the teacher in each individual case. The main use of these exercises is for progress tests, for stimulus to learn vocabulary on a regular basis, for training purposes before tests and for achievement assessment during and in the end of the course.

The focus of the implemented software has been made on its functionality and the contents of the exercises rather than on the way the exercise items can be presented.

To summarize it, gapped sentences, multiple-choice sentences and a number of other exercise types and tests are considered to be useful vocabulary items for training and assessing learner's vocabulary. The manual construction of such items, however, is a time-consuming procedure. I hope that the program that has been implemented in the course of this work and described in this essay can substitute lengthy manual construction of the learning material by automatically generating tests and vocabulary-training exercises for Swedish.

2. ICALL for Swedish: overview

2.1 CALL - overview of development

CALL – computer assisted language learning – is the area of pedagogy and technology concerned with computer applications designed for language learning. CALL era started with strong enthusiasm, it was believed that CALL would have revolutionary power over language teaching.

Originally CALL programs were collections of simple rigidly-controlled “drill-and-kill” exercises in grammar and vocabulary. As computer technologies developed, more complex programs could be designed, e.g. supporting activities for reading, listening, grammar and vocabulary training. Exercise creation of such “drill-and-kill” items has always been manual and labour-intensive, however their availability has made computer-delivered materials feasible. Their strong drawback consists in the fact that they are predetermined in all choices, e.g. pre-selected content material or rigid learning path through the program (Ramsden 2002).

Later multimedia and graphics have become a part of CALL, making learning materials more attractive. Among them I can name (interactive) exercises, instructional games, simulations and audio-video-based materials delivered on CD-ROMs. Those materials have been criticized by some researchers and end users for being flashy and not necessarily functional or error free. It is said that products with more features have higher risk to malfunction (Meskill, Anthony, Hilliker-Vanstrander, Tseng & You 2006).

Web-based materials have appeared when Internet has gained popularity and there appeared an initiative to store items in item banks making them accessible to test constructors and other teaching personnel worldwide. Item banks have made it possible to create adaptive and thus more flexible tests. Idea behind word banks is the following: each item has to be manually created, which demands investment in time and efforts. If items created worldwide (for the same language, language training purpose and proficiency level) can be stored in the same bank, then materials can be reusable and save thousands of man-hours on item construction. Item producers have therefore been faced with a need of encoding standards. IMS Global Learning Consortium is one example of implemented guidelines for shared teaching materials.

Pre-determined character of the above-described CALL materials has been the cornerstone for a lot of language teachers who wanted to decide themselves what material should be trained in exercises, games and other computer-delivered teaching materials. That has given inspiration to creation of authoring tools. One type of authoring tools makes it possible for teachers who cannot encode exercises for web applications to create their own web-delivered materials by typing text into slots (e.g. “HotPotatoes”; a lot of learning platforms can offer this possibility, e.g. “Fronter”). Another type of authoring tools is represented by (language-independent) “exercise generators” that generate exercises by simple manipulation of a text, like scrambling the order of sentences or making gap cloze exercises by removing every n^{th} word, i.e. without analyzing the input text or not taking into account word class information. Advantage of the first type of the authoring tools is that the user has influence over the contents of the material and the

items can ideally be stored in item banks; disadvantage is that it is time-consuming to produce learning materials with authoring tools like “HotPotatoes”. In the case of “exercise generators” it is the lack of linguistic analysis that makes exercises too simple in nature and allows too little user influence over the content of the exercises (except that the input text is selected by a user).

There is no denying, however, that all of the above-mentioned technologies, when used appropriately and to the task, are highly applicable. Computerized materials in language teaching do not necessarily increase efficiency of language learning unless they have the necessary functionality. In the end it does not matter whether computer tools and materials are simple or advanced: they are valuable if they are applied appropriately.

Some researches make claims that CALL has not lived up to its promises (Laurillard 2002; Ramsden 2002; Meskill et al. 2006). One of the reasons for that is said to be the fact that the development of CALL has been driven by the potential of technology rather than pedagogy and has therefore been criticized by teachers. Another reason that is named in connection with this is the technological determinism of CALL programs that takes no account of individual needs of a learner. Software developers seem to assume that they know better about how a student should learn and therefore offer a rigidly controlled studying path through the program (Ramsden 2002).

Yet another reason for CALL failure among teachers is said to be teachers themselves. They tend to use technologies to maintain their practices rather than revolutionize them (Meskill et al. 2006). Technology therefore often instead of being used (inter)actively becomes an expensive way of illustrating a lecture or class material. Unfortunately, such practices can lead to “a reinforcement of the message that education is passive reception of quantities of (entertaining) information” (Ramsden, 2002, p.160). As a result there are a lot of products on the market, expensive in production but rather underused by the target group.

Eventually the initial admiration for computer possibilities and the subsequent skepticism to CALL have been replaced by a sober and more realistic view of CALL – as a tool (among other tools) to facilitate and reinforce language learning, a complement to a teacher, rather than a surrogate “intelligent tutor”.

2.2 ICALL - overview of development

ICALL – intelligent computer-assisted language learning - is an area of implementing and deploying applications for language learning based on Natural Language (NL) Resources and Language Technologies (LT) (i.e. Natural Language Processing (NLP) otherwise called Language Engineering (LE) or Computational Linguistics (CL)) (Borin 2002b). In other words - ICALL applications are based on language-specific analysis tools that can analyze language samples (text, speech, words, etc.) and have generative power of applying the same analysis model to different language samples again and again, being an infinite source of language “wisdom” (e.g. automatic error corrections, automatic exercise generators, etc).

It has been underlined many times that language learning community, including CALL implementers, have neglected the development within NLP. At the same time ICALL within computational linguistics has also been overlooked by computational linguists (Kempen 1996; Tufis 1996; Zock 1996; Borin 2002a; Borin & Cerratto 2002). It is a frequently mentioned fact in the ICALL community that (I)CALL has not even been mentioned once in the famous collection of articles “Survey of the State of the Art in Human Language Technology” (Cole 1997) – a collection that is claimed to provide an overview over Language Technologies and Computational Linguistics as a whole and their application areas (Kempen 1996; Borin 2002a; Borin & Cerratto 2002). A good discussion as to why the two areas seem to avoid each other is given in (Borin 2002a).

It is obvious though that ICALL holds an undeniable potential for applying NLP tools and NL resources in real-life conditions as opposed to laboratory tests and academic research. ICALL can help popularize NLP tools and NL resources among a lot of users. At the same time NLP technologies and resources can support a lot of teachers relieving them from tedious tasks that can be modeled and left over to computers.

First steps towards intelligent CALL have been taken when annotated corpora have appeared. The popularization of the use of corpora in language teaching is assigned to Tim Johns (Leech 1997), who claimed that instead of allocating too much intelligence to computers and expecting them to take over a teacher’s role, we have to realize that computers are in fact stupid and cannot replace a person in a sophisticated activity like teaching, but they allow fast information processing. We can store information into them and then effectively use it for the applicable purposes, employing computer’s speed and calculation abilities (Higgins & Johns 1984; Higgins 1995).

Before considering whether computers can aid the language learning process, we need to have a clear idea of what activities are involved in teaching and learning languages. For their speed and accuracy, computers are mere machines. They can replicate human activity – but only if the activity can be comprehensively and unambiguously described. Is teaching such an activity? (Higgins & Johns 1984) p.7

Intelligent tools for language learning are within reach given the availability of key components: corpora, lexicons, tokenizers, lemmatizers, morphological analyzers, parsers etc. (Nerbonne & Smit 1996; Tufis 1996). Depending upon the aim of the ICALL application the above-named key software can be assembled in various ways making use of their different features, thus facilitating diverse learning aims. Further refinements can be added to ICALL applications given the availability of more complex tools and

resources, e.g. semantic disambiguators, keyword analyzers, learner corpora, dialogue techniques etc. It might sound as an (educational) assembly line, but the fact remains: already existing resources and tools can be successfully reused and combined as modules into pedagogically functional applications.

Nowadays various ICALL applications can support reading and writing activities, vocabulary, grammar and even pronunciation and listening training. I will in short exemplify some of those areas by mentioning several representative ICALL applications (not for Swedish, however. Applications for Swedish are described in subsection 2.5).

- REAP is a system that supports reading development and text-based vocabulary training. It first creates a student model – passive and active models, where student vocabulary knowledge is the decisive factor (Brown & Eskenazi 2004). The system then searches web-resources for texts that match learner abilities, using learner vocabulary knowledge as a primary indicator of his/her language ability. Learner levels are identified according to a 12-level scale used in language curriculum, with statistical language models built to represent each level.

A number of filters are used to ensure that appropriate texts are selected. First, web texts are parsed so that only documents containing well-formed sentences are selected, whereas those that contain lists (e.g. menus) are ignored. Second, documents are analysed for their lexical and grammatical structures to obtain their readability index. The readability index informs the grade this text can be assigned to (Collins-Thompson & Callan 2007; Heilman, Collins-Thompson, Callan & Eskenazi 2007). Third, texts are selected according to presence of target vocabulary and student's interest areas. Target words are marked in the text (Heilman, Collins-Thompson, Callan & Eskenazi 2006). Unknown words can be looked up in a companion dictionary that comes with the system. Every look-up is traced by the system and can later be used to identify difficult vocabulary and to enrich student's profile (model).

Once the text is read, a number of exercises for vocabulary training are automatically generated. Among those are definition exercises⁴, synonym exercises, cloze exercises, wordbank items, multiple-choice items, etc. (Brown et al. 2005). It is in the near future that the authors plan to extend the system with grammar training exercises and free response exercises (exercises based on free writing).

REAP has thus a generative power and desirable adaptivity to facilitate individual approach to training reading and vocabulary. A number of NLP tools and techniques are used in the system: lexicon, statistical level models, syntactic parser based on probabilistic context-free grammar, WordNet for exercise generation (Brown et al. 2005; Heilman & Eskenazi 2006). POS tagging of selected texts and semantic disambiguation are planned in the near future (Heilman et al. 2006).

- GLOSSER is another system for training reading. It is aimed at Dutch learners of French. Once a text is pasted into the application window, its every word undergoes morphological analysis and a dictionary entry for the word is recovered. When a textword is selected with a mouse the word, its morphological analysis, its meanings and examples from corpus appear in the window near the text. Words that can potentially have several

⁴ Different types of exercises are described in chapter 4.

possible morphological analyses are disambiguated with POS-tagging before morphological parsing. POS-tagger, morphological analysis software, online dictionary and annotated corpus are used in this system making it a robust ICALL application (Nerbonne & Smit 1996).

Another type of software designed to support reading is used in applications that generate TEXT-BASED CONTENT QUESTIONS. Some examples of such systems have been described in section 1.2.

- The CRITERION Online Essay Evaluation Service is a web-based application for automated essay assessment, supporting of writing process and automated feedback generation. The system consists of the two auxiliary components: e-rater that handles essay assessment and Critique Writing Analysis Tool that analyses input text, generates feedback on the basis of the analysis and thus provides necessary support in the process of writing. The system has been implemented with the intention to relieve the teacher, yet not to substitute the teacher. Teacher has a control over the tasks and a possibility to add his/her feedback or change the mark that the system offers.

Automatic essay grading has been shown to assign approximately the same grade as a human grader would assign (Monaghan & Bridgeman 2005). E-rater is based on a corpus approach and analysis of sample essays. Approximately 200-300 manually-scored essays on a given topic are necessary to build a model of an essay corresponding to a certain grade on a 6-grade scale. E-rater consists of a syntactic, discourse and topical-analysis modules. A syntactic parser is used to identify certain grammatical structures that are considered of importance (e.g. subjunctive mood, subordinate clauses). Analysis of discourse markers (e.g. first, second, perhaps, in conclusion, etc.) is used to evaluate discourse structure and analysis of vocabulary (word vectors) is used for assessing topical content. Assumption is that a good essay will resemble another good essay from a corpus of essays.

Critique Writing Analysis Tool has modules that allow it to analyze text and identify errors of different kinds: grammar, usage, mechanical errors (e.g. spelling), stylistic errors, etc. These are used in generating feedback and recommendations on how to improve the essay. System is trained on a large corpus annotated for errors. The system extracts bigrams and counts their frequencies. The bigrams that are less frequent are assumed to be errors (Burstein, Chodorow & Leacock 2003).

Criterion is a perfect example of NLP tools in service of language teaching. Tools and techniques from various areas of Computational Linguistics are used in the system.

- Other automated WRITING SUPPORT TOOLS are described and evaluated in the two overviews of automated essay assessment systems (Valenti, Neri & Cucchiarelli 2003; Dikli 2006) as well as in some descriptions of systems that are not mentioned in the overviews (Foltz, Gilliam & Kendall 2000; Kintsch, Steinhart, Stahl & Group 2000; Riedel, Dexter, Scharber & Doering 2005; Williams & Dreher 2005). Some systems are aimed at essay writing and automated assessment of essays, others at supporting summary-writing based on a provided text.

- Different needs can arise that are specific for language training, e.g. AUTOMATIC SCORING OF FREE ANSWERS. The area is vast and different techniques can be used

to evaluate short responses/free writing. Some of the techniques are described in (Burstein, Wolff & Chi 1999; Collins-Thompson & Callan 2007).

- FEEDBACK GENERATION for ITS (intelligent tutoring systems) is an important component that is present in almost all more or less complete language training systems. Some examples of feedback generating systems are described in (Nagata 1997; Haller & Eugenio 2003; Eugenio, Fossati, Yu, Haller & Glass 2005; Lu 2006) for non-language learning purposes and in (Ammerlaan 2002; Riedel et al. 2005) for essay writing training.
- Computer-supported PRONUNCIATION TRAINING has started to gain grounds in educational settings, see for example project FLUENCY described in (Eskenazi 1999).
- DIALOGUE-BASED intelligent tutoring systems and AI-BASED EDUCATIONAL GAMES do not seem to dominate the ICALL area so far, most probably because dialogue techniques are yet in an experimental phase and are used mainly for laboratory experiments. Yet some attempts are being taken (Dorr, Hendler, Blanksteen & Migdalof 1993; Johnson, Vilhjalmsson & Marsella 2005; Jung Hee, Freedman, Glass & Evens 2006).

Use of ICALL applications in real-life language classroom has been tested and documented in a number of articles, showing positive responses from both teacher and student sides, and demonstrating positive effect on learning outcome and time effectiveness (Mitkov & Ha 2003; Monaghan & Bridgeman 2005; Heilman et al. 2006).

As becomes clear from the short overview of ICALL applications given above, ICALL is a vast area where NLP technologies can make difference. Lexicons, corpora and other NL resources constitute the obligatory part of ICALL applications. In certain cases lexicons and corpora need to be specifically designed and built for the ICALL application in mind. Advanced NLP tools and techniques are used to inform ICALL applications necessary functionality. Those tools and techniques cover almost all areas of Computational Linguistics, i.e. text extraction, speech recognition, spoken language understanding, syntactic and morphological parsing, semantic disambiguation, statistic language modelling, summarization and many others. It wouldn't be too bold to say that any of the possible NLP tools and technologies can be adapted to the purposes of language learning.

2.3 Swedish as a Second/Foreign Language

The area of Swedish as a Second Language include the following related, yet different, areas of human activities:

- teaching of Swedish for non-Swedish speakers,
- assessment of Swedish for non-Swedish speakers (recognized tests in Swedish for immigrants),
- research within the area of Swedish as a Second Language,
- development of materials and computer applications for learners of Swedish,
- maybe even teacher-training program in this subject.

Below follows a short introduction into some of the above-mentioned perspectives, just to introduce the reader into the complexity of this subject.

2.3.1 Teaching/Testing Swedish as a Second Language

A number of universities and schools in Sweden offer courses in Swedish as a Second or Foreign Language, to be more particular, 11 Swedish universities out of 17 (including net university) and 5 Swedish university colleges (swe. högskola) out of 23 (collected information from www.studyinsweden.se and individual sites of each university and university college).

Among other providers of courses in Swedish there are schools offering SFI (Swedish for Immigrants) courses supported by state and free for immigrants; a number of commercial schools that offer both courses of general Swedish and Swedish for specific purposes (e.g. ABF, Folkuniversitetet, Företagsuniversitet, Lernia, Medborgarskolan); a number of e-learning alternatives (e.g. <http://www.liberhermods.se/>, learnsweden.com, eBerlitz, Folkuniversitetet). One of the e-learning courses for Swedish learners is evaluated in (Bergström 2007).

Anyone can test his or her knowledge of Swedish using placement, diagnostic or self-assessment tests. Some available tests are:

- from Folkuniversitetet
<http://www.folkuniversitetet.se/templates/PageFrame.aspx?id=80286>
- from Medborgarskolan
<http://www.medborgarskolan.se/upload/Amnesomraden/spraktester/Svenska.pdf>
- from Lingu@Net <http://www.linguanet-europa.org/plus/en/level/tools.jsp>
- from DIALANG www.dialang.org (included even in Lingu@Net resources).

There are several recognized tests in Swedish as a Second/Foreign Language:

- TISUS - Test In Swedish for University Students - intended for people who want to study at a Swedish university and need a necessary degree to be able to qualify for the studies;
- SWEDEx - SWEDish EXamination - a test in Swedish according to Council of Europe Common European Framework (CEF) of References for Languages
- SFI - Swedish For Immigrants - the first test in Swedish that is usually offered to all non-Swedish residents in Sweden for free, including training before the test;

- Tests according to CEF – majority of courses at Folkuniversitetet are aimed at CEF levels in language skills (A1/A2; B+/B-; C1/C2 etc).
- There are even some tests and courses in Swedish for Professionals, e.g. Stockholm Chamber of Commerce Certificate in Business Swedish (<http://www.foretagsuniversitetet.se>), Swedish for Medical Staff (Folkuniversitetet), etc.

2.3.2 Research within Swedish as a Second Language. Linguistic & Pedagogical Perspectives

Research within Swedish as a Second Language is a vast area, comprising linguistic studies of learner language, pedagogical and psycholinguistic studies, and socio-linguistic studies (Borin & Cerratto 2002). Although psycholinguistic and socio-linguistic studies are very interesting I will leave them outside the present essay, and limit myself to the linguistic and pedagogical perspectives.

This particular area of research comprises research into bilingualism, acquisition of Swedish as a Second Language by grown-ups and children, translation, multilingualism, etc. (see Borin & Cerratto 2002 for more details) and their application in teaching Swedish to non-Swedish speakers.

The linguistic perspective is represented by empirical studies of the learner language, one example of such studies being research undertaken by Ulla-Britt Kotsinas (Kotsinas 2005). She collected samples of spontaneous speech from interviews with six immigrants who learnt Swedish ad hoc, i.e. never in an academic environment, and summarized communicative strategies used by them. The most interesting findings are described under the headings of avoidance strategies, substitution strategies, tendency to overuse known words extending their semantic coverage and others.

It has become increasingly popular to study learner language using learner corpora. Collecting and annotating materials for learner corpora is a very time-consuming activity, but is very rewarding afterwards for studying different features of learner language (Borin & Prütz 2004).

There exist a number of Swedish learner corpora, both of written and oral language. Examples of those are the part of the CrossCheck Learner Corpus, SVANTE – a corpus of written learner texts (Borin 2003; Lindberg & Eriksson 2004), ASU - corpus of both learner essays and learner interviews collected under the supervision of Björn Hammarberg (Hammarberg 2005), EALA – corpus of low-educated adult immigrant spoken language collected under the supervision of Jens Allwood (Borin & Cerratto 2002). Many of the existing general and learner corpora for Swedish are collected in the IT-based Collaborative Learning in Grammar system (Saxena & Borin 2002), which is a unique tool for language studies and research. New corpora and resources are continually added to the ITG system. All corpora are annotated which makes it possible to use concordance software in studying learner language and learner mistakes, for example strategies for vocabulary and grammar use when writing or speaking. Results of such studies prove to be of importance for pedagogical approaches to teaching Swedish, as well as to selection of course book materials, structuring the sequence of grammar and

target vocabulary, etc. and in general for better understanding how language acquisition process develops.

Pedagogical perspective dwells mostly on attitudes learners of Swedish develop when (not) passing exams, factors influencing learning successes and failures, influence of specific educational settings on acquisition of Swedish language, etc.

An example of bringing pedagogical and linguistic perspectives together in the same research is one of the projects in Swedish as a Second Language conducted at the University of Gothenburg. Professor Inger Lindberg and her colleagues conduct a corpus-based study of vocabulary used in course books in Swedish schools with the emphasis on vocabulary frequencies. Frequency lists are supposed to be used to train non-Swedish pupils in specific school- and subject-related vocabulary, as well as to analyze teachers' use of central and peripheral subject-related vocabulary in education. Results are planned to be used in pedagogical applications.

More about the research in Swedish as L2 see at <<http://www1.lhs.se/sfi/forskning.html>>.

As can be seen, research aims vary from purely academic (to collect empirical data about some phenomena) to practical (to apply certain findings to practice). The issue of controversy, however, is that often those working with pure research do not communicate their findings to those who may and should use those findings in practice, or vice versa. This state of affairs is often mentioned about language acquisition practitioners vs language test (or assessment) researchers (Brindley 1988; Bachman 1998; Bachman & Cohen 1998; Chapelle 1998; Shohamy 1998; Alderson 2000; Read 2000).

CALL applications for Swedish as L2

A great number of course books are available as course or self-study materials for learners and teachers of Swedish. Many of them have accompanying CDs or web-pages with texts, dialogues, exercises, tests and even reference materials like digital dictionaries, grammar reference books etc., which are a good example of CALL materials for Swedish. Yet, searching for examples of CALL and ICALL applications for Swedish can become an unpleasant experience. Internet presents too much information that is low-quality and too little information that is of use. Obviously, resources that ARE of good quality and ARE publicly available – are too difficult to find without some kind of advertisement or application PR. I take therefore liberty to recommend one valuable source of language learning materials, both CALL and ICALL in character: Lingu@Net (<http://www.linguanet-europa.org/plus/en/home.jsp>). Resources and even courses for most European languages, including Swedish, can be found there. The advantage of Lingu@Net lies in the fact that each resource found by this online service is evaluated and classified according to target language, proficiency level in target language, and source language.

2.5 ICALL applications for Swedish as L2

I have mentioned above that the Computational Linguistics community seem to neglect the area of ICALL. That is not totally true. More and more attention is being paid to this area. The obvious disadvantage, however, is that ICALL is not commercially beneficial since the majority of ICALL applications need huge resources like corpora and dictionaries which are very expensive in construction. Existing corpora cannot be used commercially due to copyright limitations, and hence ICALL applications based on such corpora cannot be commercially distributed. The dilemma is therefore where to take money to develop ICALL applications. Naturally, commercial companies are not interested in investing money into non-beneficial projects. The prevailing tendency with non-profitable funds is to give priority to projects where academic world meets industrial needs and invested money comes from the two sources – non-beneficial organizations (e.g. Scientific Council or some other governmental fund) and industry. ICALL projects that take place in Sweden are funded by governmental organizations, but the competition is very high and not many of such projects are being granted project money. There are several strong research groups in Sweden that, in spite of the financial problems, manage to get necessary funding for ICALL projects. Among them are KTH NADA group, Språkbanken group at GU, Uppsala University Learning Lab, Centre for Speech Technology (Speech, Music and Hearing Department) at KTH, IPLab at KTH and some others. Some ICALL applications for Swedish come as a side-effect of projects originally intended for other languages than Swedish, see for example VISL project below. Among commercial companies one can name Vocab AB that develops environment for vocabulary training and authoring tools for translation-based exercises; Larson Education AB that has software for training different language skills; Lingsoft that has a number of tools like grammar-checkers and spellcheckers for Swedish and WordFinder that converts

major available dictionaries into computer-readable format and develops grammar tools for proof-reading texts.

Some examples of available ICALL applications (not tools) for Swedish learners (and teachers) as well as some ongoing projects are illustrated below. Some past projects that for some reason have not resulted in publicly available applications are also mentioned. The first group is comprised of end-user products. Even though all of them are composed of a number of modules that are worth talking about separately, I am dwelling upon the systems as a whole, mentioning their functionality, as well as NLP tools and NL resources they are based on. All the applications described below are NLP systems in support of learning Swedish.

Table 1 presents an overview of the ICALL applications described below sorted according to their target group and language training purposes.

Table 1. Overview over ICALL applications for Swedish as L2

L-ge Skills/ L-ge levels	Grim	ITG	VISL	Ville	DEAL	Vocab Tool	Lingus	Word- Finder	Squir- rel	Didax	ARTUR
Writing	X										
Reading						?	X*		X		
Listening				?	?		X*				
Speaking				X	X		X*				
Grammar	X	X	X				X	X			
Vocabulary		?				X**	X*				
Pronunciation				X	?		X*				X
Testing							X*			X	
Beginner level			X	X	X	X	X	?	X		X
Intermediate level			X			X	X	?	X		
Advanced level	X	X	X			X	X	?	X		
Native Speakers/ Researchers	X	X						?	X		
(Computational) Linguistics Students	?	X	X						?		

X* non-NLP-based modules

X** translation-based exercises

2.5.1 GRIM

GRIM is a language learning environment for supporting of writing. This application is aimed at both native speakers and learners of Swedish. The user can write a text in Swedish and receive immediate feedback from the system in the form of detected spelling and grammar errors and suggestions for their correction. The system also offers some other sophisticated features like identification and highlighting of certain parts of speech, word-processing functionality, etc.

Grim consists of a number of NLP tools that are incorporated into the system (Knutsson 2005):

- Rule-based grammar checker Granska that contains tokenizer, PoS-tagger, spell-checker STAVA, a number of rules describing correct syntax as well as rules for identification of errors (Carlberger, Domeij, Kann & Knutsson 2004);
- Shallow parser GTA;
- ProbGranska – a probabilistic grammar checker;
- SnålGranska – a grammar checker based on machine learning training (Bigert, Kann, Knutsson & Sjöbergh 2005);
- Word-processing functions;
- System for generation of feedback (suggestions for correction) based on identified errors in Granska.

To make this system work properly different language resources have been used for training, deriving of wordlists, etc.:

- SUC
- Lexin dictionary
- Concordances from Parole
- Learner corpus Svante (CrossCheck)

Class experiments with Grim have shown that the system detects correctly relatively many errors and suggests correct answers. However, professional writers of Swedish find it more comfortable to work with Grim than learners. Native speakers more easily forgive incorrect error detection than learners (Knutsson, Cerratto Pargman, Severinson Eklundh & Westlund 2005). It is difficult to say how common this system is among language learners and whether it is widely spread in the language learning settings.

GRIM is freely available at <http://skrutten.nada.kth.se/grim/>.

2.5.2 IT-based Collaborative Learning in Grammar (ITG)

A useful ICALL software for Swedish and Linguistics Studies is IT-based Collaborative Learning in Grammar (ITG system). It is a corpus-based grammar tutor aimed at students of Linguistics and Computational Linguistics as well as at researchers (Borin & Dahllöf 1999; Borin & Saxena 2004). The system consists of a linguistic encyclopedia with descriptions of grammatical concepts and constructions, corpora, corpus search tool, resource module and an interactive grammar exercise module (Saxena & Borin 2002). To make the system work properly and to avoid a number of incompatibilities already existing corpora had to be converted to a uniform XML format. The system is ever-growing since more and more corpora are added to it. The same refers to the interactive exercise module. The target group for the system may also expand as the system develops to include even teachers and learners of Swedish as a Second Language.

Possibility to use any of the corpora contained in the system makes it an invaluable resource both in research, studies of languages, and even in teacher-training programs when it comes to teachers of Swedish and Swedish as a Second Language. There are a lot of corpora in the ITG system, not necessarily Swedish – other languages are also represented, e.g. an annotated corpus of Kinnauri, a corpus of lesser-known language. The corpus search tool is designed to represent not only concordance lines, but even to visualize graphically corpus structure with a possibility to get access to full

sentences/texts and to see the distribution of a search query in a corpus. ITG corpus searching tool employs a lot of ideas from ETAP-WebTEq, a corpus search tool described in (Olsson & Borin 2000).

Different exercises for grammar training are automatically generated by the system, e.g. training of syntactical functions in a sentence and identification of parts of speech (Borin & Dahllöf 1999; Borin & Saxena 2004).

The system makes use of a number of NLP tools and resources: PoS tagged and syntactically annotated corpora, parsers, corpus search tools, grammar writing tools, visual presentation of corpora maps. ITG is not publicly available, but can be used for free for academic purposes. It is extensively used for teaching Linguistics and Computational Linguistics at the University of Gothenburg, Uppsala University and Stockholm University.

ITG system is freely available for academic purposes, contact Lars Borin, <<http://spraakbanken.gu.se/personal/lars/>>

2.5.3 VISL - Visual Interactive Syntax Learning

Another example of an NLP-based computer program for language learning (for Swedish, among other languages) is VISL – Visual Interactive Syntax Learning. The system has been developed in Denmark and initially was a less ambitious project than it turned out to be in the end (Bick 2001). Instead of originally planned four languages (English, German, French and Portuguese) with application strictly at Odense University (Denmark), it now comprises materials and interactive exercises for more than 14 languages (languages are not equally “equipped”) and is used at a lot of places over the Internet (Bick 2005). Swedish is included among the available languages and a number of modules have been developed for Swedish, among them interactive exercises for training grammar and syntax. The following is available for Swedish:

- sentence analysis with a graphical representation in tree format, tagger-format and a number of other formats;
- games: labyrinths, shooting gallery, paintbox game and wordfall for part-of-speech training; and space rescue and syntris for training of syntactical features of a sentence;
- machine translation from Swedish to English and Danish.

For other languages there are included corpora, text analysis, quizzes in different language aspects and many other options. Especially well-“equipped” are Danish, English, Esperanto, French, German, Portuguese and Spanish.

Language analysis in VISL concentrates on surface structure and form-function dependency. Constraint Grammar is the core approach to analysis. VISL system is built upon corpus-based approach to exercise generation. Games are run in Java applets and are colorful and entertaining. Machine translation, spell/grammar checkers, question/answering system are among numerous NLP tools used in the system.

VISL is freely available at <<http://visl.sdu.dk/>>, with its Swedish component at <<http://beta.visl.sdu.dk/visl/sv/>>.

2.5.4 Ville & DEAL

DEAL is a dialogue system that is under development at the moment at KTH, Centre for Speech Technology (Speech, Music and Hearing Department). Its purpose is to combine dialogue techniques with language learning in a stimulating and entertaining game setting. The module that is now under development is set in a flea market where a learner can communicate with sellers of different things and thus train conversation skills (Hjalmarsson, Wik & Brusk 2007).

DEAL is a free-standing part of Ville, a framework for language learning. Ville is a dialogue agent that functions as a tutor in pronunciation training. A learner says a word through a headphone, the word is recorded and is saved into the student profile with date and time tags; the system analyzes the input and comes up with suggestions what need be improved (Bergström 2007). Ville can thus detect and correct pronunciation errors. There are several Dialog Managers (DM) built into the system to take care of conversations in different domains, so that the problem of one all-knowing Dialogue Manager is avoided. Ville offers training on the level of pronunciation – phones, syllables, words, sentences, intonation. DEAL takes this training a step further offering the learner a possibility to practice conversation. In DEAL the dialogue agent is no longer a tutor giving corrective feedback, but a conversation partner.

Ville's architecture comprises a number of DMs, pronunciation analyzer, text-to-speech module, automatic speech recognition module, teaching strategies, 3-D animated head, and a student profile module (Wik 2004).

DEAL is based on Higgins, a spoken dialogue system, employs discourse modeler Galatea, modules for semantic interpretation, chart-parser, probabilistic speech recognizer, word-chunking techniques and a number of other NLP techniques (Hjalmarsson et al. 2007).

Both Ville and DEAL are still in research phase. See more information on <<http://www.speech.kth.se/ville/>>.

2.5.5 ARTUR

ARTUR – a multi-modal ARTiculator TUtoR - is an ongoing project at KTH, IPLab. ARTUR is a system that will demonstrate to the user how to pronounce different Swedish sounds visually and acoustically; and provide speech production feedback. The system will identify articulation mistakes by analyzing the position and shape of the user's tongue from received utterance; phonological mistakes made by users through facial movements will be identified through the state-of-the-art phoneme speech recognition. Feedback given to the user will consist of whether pronunciation is accepted or not, what articulation parameters the user should concentrate on as well as visual demonstration of how to pronounce words/segments of speech.

The novelty of the approach accepted in ARTUR is that the pronunciation learning can be supported by other means of modality than hearing (Eriksson, Bälter, Engwall, Öster &

Kjellström (formerly Sidenbladh) 2005). The target group for ARTUR is second language learners, hearing-impaired persons, speech therapy patients.

Information about ARTUR as well as publications can be found on <<http://hci.csc.kth.se/projectView.jsp?name=artur>>

2.5.6 VocabTool

VocabAB is a company that develops and delivers a specially designed commercial platform for training Swedish vocabulary, VocabTool. VocabTool is based on frequency lists and is offered for three levels – V3000 (= beginner), V5000 (=intermediate), V7000 (=advanced). Learners can upload or paste any Swedish text into the application window, which is then analyzed, each word being hyperlinked to a (proactive) dictionary entry. If the word has several entries in a dictionary, it is possible to see all of them. Text vocabulary is analyzed for its appropriateness, and words are marked in the text in three colors, one color for each level. The learner can thus concentrate on the automatically identified target vocabulary.

Once the text is read, there is a possibility to train words in exercises, the latter being of two types – flashcards and “fill-in-the-gaps” items. Exercises are translation- or definition-based. Language pairs that are available in the application are Russian-Swedish, Spanish-Swedish, English-Swedish, Swedish-Swedish and German-Swedish. German-Swedish is not available for the level V7000.

The user marks the words from the glossary that he/she wants to train by ticking the boxes. Flashcard items consist in finding a translation equivalent to the word or phrase that is shown. The system selects the correct answer to the Flash-card item by consulting an appropriate dictionary. For Swedish-Swedish pair a definition of the target word/phrase is shown in the key. The application cannot correct this exercise type, leaving it to the user to compare his/her answer with the suggested answer. The user then has to decide whether his/her answer is correct or not. The user is supposed even to mark a vocabulary item as “learned”. Items that are not yet “learned” will automatically appear again and again during the study process.

“Fill-in-the-gaps” items are, too, based on translations. The user is supposed to write a Swedish equivalent of the omitted word. This exercise can be automatically corrected.

VocabTool has not published any articles that would allow judging to which extent NLP technologies have been used in producing learner materials. The only fact known to me from communication with Lars Borin is that text processing is based on the morphological mechanism MoWa (Niwinski 2002). Furthermore, the trial version bought on vocab.com page suggests that the LEXIN lexicons available online are used in the software. When compiling a glossary certain words are not included, most probably those that are usually classified as stop-words: *det*, *att*, *deras*, *detta*, *därför*, etc. Some other words that are rather infrequent in character also seem to be overlooked by the program, i.e. no dictionary entry is created for them. There can be different reasons for ignoring them – first, these words might not be considered to belong to any of the three learner levels; second, the online dictionary might lack entries for these words. If LEXIN

lexicons are used, the second reason is the most probable, since LEXIN dictionaries contain approximately 20 000 words each. Another fact that I have observed while using the trial version is that words that have several entries in a lexicon are not disambiguated, i.e. all possible entries – including entries for different parts of speech - are offered for training. This suggests that no part-of-speech tagging is done to text words.

VocabTool is a commercial application, see <<http://www.vocab.se/>>

2.5.7 Lingus

Lingus is a combination of both CALL and ICALL modules blended together in a system aimed at learners of different languages, among others Swedish. The system offers ready-made exercises in grammar, pronunciation, vocabulary as well as authoring tools for grammar, vocabulary, listening comprehension, reading comprehension, spelling, and pronunciation. One of the authoring tools, GramLing, is designed for creating grammar exercises and is based on NLP technologies, namely morphological analyzer, see (Olausson Källfelt & Fogelberg 2004) for details. The NLP-based module has been created by Wojtek Niwinski on the basis of another program, “CALLe svenska”. Both programs are based on a morphological mechanism MoWa (Niwinski 2002).

Integrated speech analysis tool in GramLing allows intonation curve analysis. The system collects learner statistics and creates a learner profile. It is claimed that the learner environment is so modular that it is possible to deploy any NLP tools, e.g. speech analysis tools, to inform better functionality to the system.

Lingus is a commercial software, see <<http://www.larsoneducation.se/>>

2.5.8 Wordfinder

Wordfinder offers a number of computerized dictionaries with a search engine, that however seems to look up words according to the way they look in a text (graphical form) or given in query, without prior lemmatization. The company offers even a package of tools to support Swedish writers of English “Wordfinder Proofing Tools” as well as grammar support for writers of Swedish “Skriv Rätt”.

Wordfinder is a commercial application, see <<http://www.wordfinder.se/>>

The applications described above are in active use at the moment of writing.

There have been a number of very interesting and promising initiatives within (I)CALL for Swedish. It seems, however, that as soon as project money is exhausted, those initiatives are abandoned or for some reason are made unavailable over the Internet. The interested readers/users have to content themselves with project reports, magazine and conference articles and grieve over the absence of what seems to have been so near to actual use. Among such “imagination-teasers” I can name Squirrel project (Borin 2002a; Nilsson & Borin 2002; Carlson, Grönroos & Lemmilä 2005) and Didax project (Babic 2002; Bengtsson & Lingdell 2002).

2.5.9 Squirrel

Squirrel is an ICALL project run by several Nordic countries aimed at creating a prototype Internet browser for teachers of Nordic Languages with a functionality of automatic locating and extracting authentic learner texts from the Internet according to language, topic and difficulty level. The tool can analyze texts in several Nordic languages, including Swedish, and find texts that are similar in topic/key vocabulary to the one submitted by the user (Nilsson & Borin 2002). The system contains modules for extracting key words (query terms) from example document submitted by the user, html-parser, a module for automatic language identification when searching for relevant texts on the Internet, word tokenizer, stemming module and readability analysis module (Nilsson 2003).

In 2005 the project was still active, at least in some of the Nordic countries (Carlson et al. 2005).

2.5.10 Didax

Didax is an example of CALL representing a system for online testing. In short, it is a combination of authoring tool for teachers and test environment for students. Teachers can create different test items and combine them into a test. There is an authoring tool for multiple choice and fill-in-blank questions. Test items are stored in a QTI format, which is the biggest advantage of the system compared to the majority of other learning platforms that were available at the time when Didax project was in progress. Students get access to their profiles, do tests, get automatic feedback in terms of right/wrong; teachers log in and grade students' tests (Babic 2002; Bengtsson & Lingdell 2002). By description of it the system is language free, i.e. can be used for any other languages than Swedish. No NLP elements were included into the system at the time of publication, but the ambition to later incorporate intelligent NLP-based modules was present (Borin, Åkerman Sarkisian & Bengtsson 2001).

2.5.11 Other projects

Some other applications and resources both for Swedish language and other languages, especially those applicable to grammar learning, are described in an overview by Hammarström (2002). Another project that is worth mentioning is Scribani Project, where writing tools and environment for collaborative writing are developed (see <http://www.nada.kth.se/iplab/scribani/>), <http://hci.csc.kth.se/projectView.jsp?name=scribani>).

It is useless to speculate about why there are so few end-user ICALL applications for Swedish. But several reasons are obvious: as has been mentioned above, implementing NLP-based systems for language learning is expensive and requires money. Since most of ICALL products need to have access to Swedish corpora – and those have copyright limitations for commercial distribution – these systems cannot be later commercially

distributed. Funding has therefore to come from non-commercial organizations, which means tough competition with other project applications. Other reasons are very well described in (Borin 2002a), in brief – teachers and software programmers belong to different cultures and therefore misunderstand each other; technology does not live up to demands of communicative pedagogy; teachers are in general technophobes and therefore avoid computers in teaching, whereas software developers assume that they know better what problems to address and offer solutions to non-existing problems instead of addressing existing ones; it is also often a fact that software developers assume to know what and how students should learn and therefore develop software that is not accepted by teachers and/or students at schools.

2.6 NL resources and NLP tools for Swedish

The above-described applications for Swedish are examples of direct use of NLP tools in service of language learning. There is, however, even need for indirect use of NL resources and NLP tools in ICALL, examples of which are taggers that can tag learner corpora, and learner corpora that can assist in error identification in written texts.

The key idea for this work is to analyze feasibility of development ICALL software for vocabulary training reusing publicly or academically available resources and tools, adapting to the current demands in standards. It is therefore important to be aware of what is available for Swedish today. Unfortunately, it is easier said than done. It is next to impossible to make a complete list of existing resources and tools within the time limits for a Master Thesis.

In the next chapters I will analyze one NL resource that I have used in my application, namely, SUC – Stockholm Umeå Corpus. Corpora and lexicons, above all other resources and tools, contribute to integration of technology into pedagogically valuable and practical applications. They are the “core knowledge” that computer programs are bestowed with and are necessary when analyzing language production and generating output for the learner.

3. Use of Corpus in the Exercise Generator

The most important language resource used in this generator is Stockholm Umeå Corpus (SUC). It is used in several modules (GapCloze Items, Multiple Choice Items, Wordbank Items, and Total Vocabulary Test). SUC is a major unique source of authentic examples, sentences and texts for vocabulary item generation. It has also provided the frequency information for classifying vocabulary into frequency bands (eight bands) that are essential for all the modules of the implemented exercise generator.

In the light of the above-mentioned, I consider it appropriate to touch upon corpora in general, give a concise overview over corpora for Swedish and to describe SUC in detail highlighting its benefits and disadvantages from two perspectives: that of a computational linguist and that of a teacher of Swedish as a Second Language. Both perspectives need to be blended together in an effort to build a pedagogically useful application for computer-assisted generation of vocabulary exercises.

3.1 General on corpora in Second Language Acquisition

Corpora have been used for a long time for research and of late have gained popularity even in teaching languages. Study of the literature on corpora shows that corpora within language teaching is mainly used in the form of concordances and frequency lists (Gavioli 1997; Leech 1997; Minugh 1997; Hunston 2002; O'Keeffe & Farr 2003; O'Keeffe, McCarthy & Carter 2007). Corpora can, however, offer a lot more than that for language learners and teachers. Leech (1997, p.1) for example mentions a “whole range of largely unpublicized pedagogical activities making use of corpora”.

Corpora offer a rich resource of authentic data, grammatical patterns and language features. The latter include lexical, grammatical, morphological features, collocation patterns, semantic features etc. depending upon what linguistic parameters have been annotated in the corpus. Corpora are thus a source of available and carefully encoded language information. Corpora are largely developed within Computational Linguistics, whereas the main areas of corpora application and usage are within Linguistics, Computational Linguistics, and as mentioned above, within language teaching.

The advantage of using corpora within language teaching in combination with programming skills is that the learning materials can be customized to the individual needs of learners, courses or syllabus requirements; materials can be reusable independent of time and place, automatic generation of teaching materials can save teachers' time on both production and correction of assessment items.

Anyone who has been involved in teaching languages can confirm that there is always need for new materials. Wilson (1997, p.117) for instance writes about the problem of addressing students of different levels and creating materials for that:

In language course design there are two major problems:

- (a) How to provide a range of materials to meet the needs of students with different abilities.

- (b) How to provide at every ability level enough exercises to ensure that a student is confronted by a different set of examples whenever he or she uses the language-learning program.

Corpora can address both problems in an effective way. Corpora can be applied in different ways to teaching:

- directly by using concordancing as reference to check how a particular word is used or by selecting authentic examples before the lesson;
- indirectly by extracting frequency information from corpora and using it for identifying lexical items and grammatical structures of the most importance for learners;
- or by using corpora as a source of teaching material exploiting tagged features for further (manual or automated) processing or analysis, e.g. grammar, morphology, vocabulary, semantics, collocations etc. If automated way is used for producing corpus-based teaching material, there might arise a need of disambiguation, proof-reading and correction before the item is approved for learner usage; another restriction is that scoring of computer-delivered vocabulary items has to be strictly controllable and defined in terms of correct – incorrect (unless sophisticated tools for scoring of free responses are available).

The direct use of corpora (i.e. concordancing) is a wide practice nowadays judging from the literature (Dodd 1997; Gavioli 1997; Leech 1997; Mindt 1997; Minugh 1997; McEnery & Wilson 2001; Hunston 2002; O’Keeffe & Farr 2003; O’Keeffe et al. 2007). Most teachers who practice using corpora in classroom exploit even language statistics and frequency lists. Using corpora as a source for automatically generated exercises, however obvious it might seem, is not as often mentioned in the literature but certainly is no novelty (Coniam 1997; Wilson 1997; Borin & Dahllöf 1999). The second and the third application of corpora mentioned above will be discussed later in this chapter.

3.2 Overview of Swedish corpora

However obvious the meaning of the term “linguistic corpus” might seem at the first glance, corpora is understood differently by different people. Meyer (Meyer 2002) describes a posting on one of the corpora forums where the sender was wondering where (s)he could find an online corpus of proverbs. This message sparked a heated discussion about what a corpus is. Is computerized collection of proverbs a corpus? Is an online dictionary a corpus, too?

Questions are many. Should an online library be defined as a corpus? How to treat a computerized newspaper archive? Should a corpus follow some design standards, have some search instruments available and be annotated in some way? According to Meyer, the answer to such questions depends on how broadly one wishes to define a corpus. Potentially, corpora can be constituted by any text type, be it raw texts, annotated texts, lexicons or word lists.

In Appendix 1 I am listing some text collections that are called corpora by other, more respectable and experienced linguists and computational linguists. The list of corpora is a result of blending together lists derived from <<http://spraakbanken.gu.se/>>, <<http://sprakteknologi.se/resources/data-collections>>, ITG system, <<http://www.ling.gu.se/projekt/tal/>>, and references to Swedish corpora I have come across in articles. Swedish specialized (e.g. terminology) corpora and parallel corpora are listed among corpora of written Swedish. The list is non-exhaustive and includes only some freely available corpora and corpora available for academic research.

3.3 General on SUC and its role in the exercise generator

Among the available corpora for Swedish, SUC (Stockholm Umeå Corpus) and Parole are the two well-annotated corpora of written Swedish that are often used for research purposes. SUC has been chosen as a major corpus for SCORVEX since it has a number of advantages. First of all, it is a richly annotated corpus. Second, it is a balanced corpus comprising texts from different genres. Third, it is representative of modern Swedish. It is the combination of these three characteristics (annotation, balance, representativeness) that makes it so valuable for applications like the one described in this work. Parole, on the other hand, even though a much larger corpus (19.4 mln. words), contains texts that cover the period of 1976-1997 and does not meet the requirement of balance and representativeness. Moreover, SUC PoS-annotation has been manually proofread and represent therefore a high degree of reliability whereas Parole has never been manually controlled and therefore cannot boast the same degree of reliability.

SUC is a collection of annotated texts in Swedish dating from 1990-s. Texts have been carefully selected to present samples of general-purpose (published) language comprising 1,2 million running words. It is said that SUC is the only corpus which is representative of modern general-purpose Swedish. It contains texts from 9 major genres and 48 domains (not including spoken language, though). Genres are represented by:

- Press: Reports
- Press: Editorials
- Press: Reviews
- Skills and Hobbies
- Popular Lore
- Biographies, Essays
- Miscellaneous
- Learned and Scientific Writing
- Imaginative prose

Each of the genres falls into a number of domains, each domain containing a number of texts. Texts have been selected and structured in such a way that allows for parallel comparative studies between SUC, Lancaster Oslo Corpus and Brown Corpus, i.e. between Swedish, British English and American English (Källgren, Gustafson-Capková & Hartmann 2006).

The tagging system consists of 22 part-of-speech (PoS) tags plus morphosyntactic tags where applicable and 4 delimiter tags (for punctuation marks). Each text word is accompanied by its uninflected form, which in combination with PoS provides its lemma⁵. Each text has a name and a genre label, and is kept in its own file, comprising roughly 2000 running words (tokens). Shorter texts are either grouped together (up to 2000 words) in one file or in certain cases are stored in separate files in spite of their small size (Källgren et al. 2006).

SUC texts have been semi-automatically tagged with a tagger developed by Lingsoft predecessor; all the texts were afterwards manually proof-read. SUC is conformant with TEI, Text Encoding Initiative, which provides general guidelines for encoding texts based on SGML and provides a standard definition of the markup, both textual and linguistic, for corpora. Of late XML format is used as well, which is SGML-conformant. SUC also follows CES (Corpus Encoding Standard). Any corpus that is CES-conformant is also TEI-conformant and SGML-conformant (McEnery & Wilson 2001). In this generator an xml-version of SUC-files with PAROLE tags is used.

SUC is a linguistic resource. The term *linguistic resource* (LR) refers to large collections of machine readable data that presuppose use of software for collection, preparation and management of data, the software being also covered by the term LR. LR are used in building and evaluating of NLP tools and algorithms (Godfrey & Zampolli 1997). SUC is both a product of computational linguistics and a computational linguistic resource with extensive annotation; therefore any application built upon it using its annotated features in an automated way is a computational linguistic application.

Corpora annotation is the key to its value as a source of linguistic information in language studies. However, having the annotation present in a corpus is one thing, using it is quite different. Below I will try to demonstrate how different linguistic features present in SUC have been applied to the tasks of this generator.

The tags that have proven to be especially significant in designing algorithms for finding exercise material in SUC include:

- text word tags for lexical searches
- uninflected forms for base form searches
- part-of-speech tags with morphosyntactic information
- sentence start and sentence end markings
- text start and end markings
- text domain labels

An example of an annotated sentence from SUC is shown below in Figure 1.

- `<s id=aa01a-004>` and `</s>` stand for start and end of a sentence plus id number of each sentence. These tags are used when selecting a sentence containing a target vocabulary item. Sentence ids have proven to be particularly useful since they contain both reference to the text file and to the running number of the sentence.
- `<w lem=.....</w>` is a headword tag for lexical searches;

⁵ More about the concept "lemma" and its use in this work see in 3.5.2

- 'avspänning' msd='NCUSN@DS' stands for part of speech annotation, used when searching for a particular part of speech, often in combination with the base form of a word; plus more detailed morphosyntactic information; useful in e.g. search for distractors for multiple-choice items;
- lem='avspänning' represents uninflected form of the word, which in combination with POS represents the lemma of the text word;

```

<s id=aa01a-004>
<w lem='avspänning' msd='NCUSN@DS' n=12>Avspänningen</w>
<w lem='mellan' msd='SPS' n=13>mellan</w>
<w lem='stormaktsblock' msd='NCNPN@DS' n=14>stormaktsblocken</w>
<w lem='och' msd='CCS' n=15>och</w>
<w lem='nedrustningssträvande' msd='NCNPN@IS' n=16>nedrustningssträvanden</w>
<w lem='i' msd='SPS' n=17>i</w>
<name type=place>
<w lem='Europa' msd='NP00N@0S' n=18>Europa</w>
</name>
<w lem='ha' msd='V@IPAS' n=19>har</w>
<w lem='inte' msd='RG0S' n=20>inte</w>
<w lem='mycken' msd='AQPNSNIS' n=21>mycket</w>
<w lem='motsvarighet' msd='NCUSN@IS' n=22>motsvarighet</w>
<w lem='i' msd='SPS' n=23>i</w>
<name type=place>
<w lem='Mellanöstern' msd='NP00N@0S' n=24>Mellanöstern</w>
</name>
<c lem='.' msd='FE' n=25>.</c>
</s>

```

Figure 1. Excerpt from SUC. An example of an annotated sentence

The way the corpus information has been used in this application can be presented by the following diagram (inspired by Mindt (1997)):

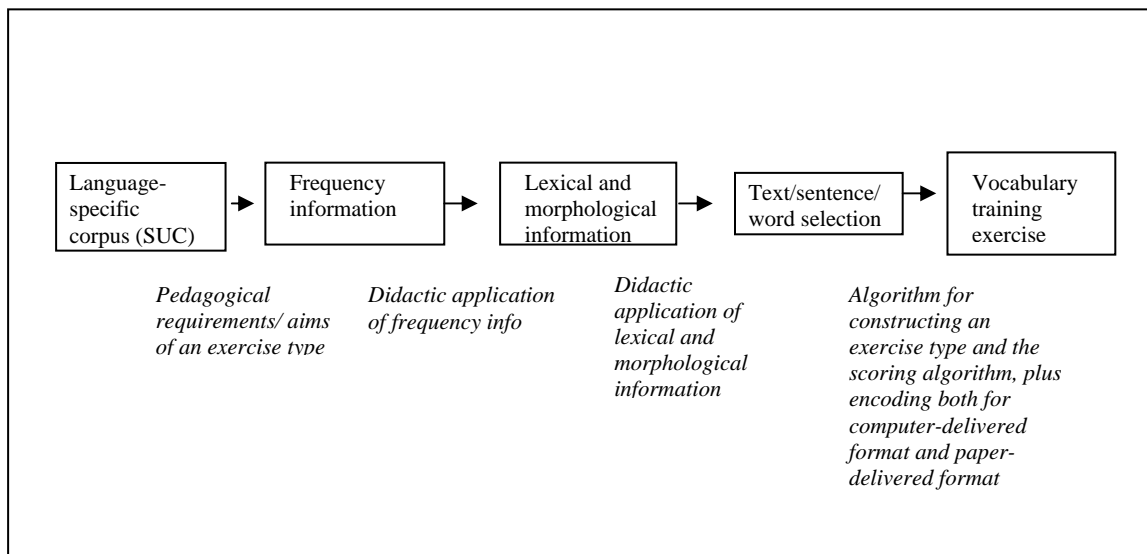


Figure 2. Schematic representation of corpus use in the exercise generator

Each exercise type (or item type) requires different language information and uses different approach to vocabulary as a construct (Chapelle 1998), which in turn addresses the language learner to different linguistic aspects of a word. When those are defined, frequency lists are consulted and necessary lexical elements are identified on the basis of morphological and/or syntactical information.

SUC serves as a source of classroom and assessment material that can be adapted to the pedagogical situation. Three operations can be performed on the corpora material for the purposes of our exercise generator:

- lexical items can be extracted in isolation through search for a lemma and its particular morphological form (e.g. distractors for multiple-choice items or target words for yes/no items);
- lexical items can be extracted in the context of a sentence for shorter exemplification of its use or for testing it (e.g. sentences with target word(s) for multiple-choice items or wordbank items);
- lexical items can also be extracted in the context of a larger text, or rather a text first is extracted and then lexical items corresponding to the level of the learner are identified for training/assessment.

One of the above-mentioned operations is selected depending upon the item type and approach to vocabulary training/assessment. The extracted elements are combined into an exercise according to the algorithm and scoring procedures are encoded.

3.4 Some words on the notions of “word” and “lemma”

The way researchers operationalize the construct “word” influences the way word statistics and frequency counts are collected and the way different aspects of individual words are analyzed. This has a direct impact upon the pedagogical application of the collected statistics (Gardner 2007). As has been mentioned above, the frequency count that SCORVEX is based upon is calculated upon lemmas. Lemma is a useful concept for applied corpus studies, but it contains a number of drawbacks. There exist different ways to define the notion of lemma. The way it is understood in SUC (and consequently the way it has been inherited by SCORVEX application) does not exactly reflect the way we would like to define it.

In SUC context lemma is understood as a set of word forms having the same stem or base form and belonging to the same word class, e.g. all occurrences of the word forms *flicka*, *flickas*, *flickan*, *flickans*, *flickor*, *flickors*, *flickorna*, *flickornas* are counted together since they have the same base form *flicka* (Eng. girl) and the same word class *noun*. This is reasonable. However, such definition of a lemma allows grouping together words that share the same base form and word class, but not grammatical features (inflectional morphological aspects), e.g. *val* (noun, -et; the neuter gender, 6th declension; Eng. *election*; *choice*) and *val* (noun, -en, -ar; the uter gender, 2nd declension; Eng. *whale*) are counted together in frequency statistics. The missing information about the declension of

a noun or conjugation group of a verb results in a partially misleading frequency information. The verb *vara* irrespective of which one of the two verbs is meant – *to be* or *to last* – has always the same frequency value, in spite of the fact that the two verbs are conjugated differently, one being a strong verb (conjugation group 4), the other being a weak verb (conjugation group 1); they also have unrelated meanings, the meaning “*to last*” being much more rarely used.

Furthermore, multiword items are not identified as units, but are rather split into constituent parts and each part is counted separately. There are some exceptions to this general approach, e.g. *bland annat* (Eng. *among other things*) are counted as one unit and not as two separate lexical items *bland* and *annat* (Eng. *among, other*).

Another aspect that is missing in SUC is derivational morphology, i.e. mark-up of root morphemes and word-building affixes of each lexical item. The suggested markup could have allowed collecting frequency statistics according to the word family principle, i.e. words that share the same root being grouped together (e.g. *lära*, *v* and *lärare*, *n* would make the same entry). The frequency statistics collected from SUC at present does not allow to group words on this principle, which means a learner that knows the verb *läsa* (Eng. *read*) cannot be not assumed to know the noun *läsare* (Eng. *reader*).

However, errors in frequency calculations of the type “*vara*, *verb* (Eng. *to be*) – *vara*, *verb* (Eng. *to last*)”, though being a systematic drawback, influence only a few rare cases in Swedish and thus have to be neglected in want of a better analysis software. Multiword items that are most frequent in Swedish are marked up as units and do not add misleading information to the statistics used for L2 learners, e.g. adverb *till exempel* (Eng. *for example*) is taken care of in the following way:

```
221 till_exempel.RG 445.332621 9 t_ex.RG0A.488
    t.ex..RG0A.113 tex.RG0A.5 t.ex.RG0A.3
```

Finally, taking derivational morphology into account is an arguable demand. Some researchers build their word frequencies upon the notion of word families but they aren't many (Gardner 2007). Thus the two features - having less frequent multiword units marked up as units and having roots and affixes marked up for each lemma - refer rather to desirable than to absolutely necessary features. Therefore, we consider word frequency statistics based on SUC the most reliable and the most appropriate one for language learning purposes available at present.

3.5 SUC as a source of frequency information

In pre-corpora times language teaching materials have been selected based on the intuition of course-book writers and/or teachers. Now that corpora are available it is possible to check those intuitions by consulting automatically generated frequency lists over different features tagged in a corpus and make conclusions about which features are most typical, e.g. most frequent and presumably most important for language learners. Some teacher intuitions referred to above can be confirmed right, others – proved wrong. For instance some language teachers working with corpora have come to an insight that certain language course books tend to overestimate importance of the verbs “will” and

“shall” as expressions of future in English overlooking the fact that native speakers prioritize other ways of expressing future.

It is also true that frequency alone cannot be the only factor for consideration when it comes to learner material selection. For example frequency statistics shows that weekdays “Tuesday” and “Wednesday” are less frequent than other weekdays. It would be irrelevant, though, to learn frequent weekdays in the beginning leaving the two “infrequent” weekdays for later training. As O’Keeffe et.al. (2007) put it, “pedagogical decisions may override these awkward but fascinating statistics” (p.41).

Nevertheless, in spite of all imperfections of the equation: ‘most frequent’ = ‘most important to learn’ (Leech, 1997, p.16), it is difficult to deny the value of the frequency statistics for selection of leaning materials. It certainly helps separate wheat from the chaff – rare examples and words should be left out for later training (McEnery & Wilson 2001).

In this exercise generator extensive use of frequency statistics over Swedish vocabulary is made. It provides ground for well-balanced frequency information not biased towards any special area of knowledge (e.g. law or medicine).

For the purposes of test item generation already existing base vocabulary pool derived from the SUC by Eva Forsbom (2006) has been used. The advantages of using base vocabulary pool are numerous. Apart from the fact that it is publicly available in electronic form from <<http://stp.lingfil.uu.se/~evafo/resources/basevocpool/>> (under the heading “Files”, data -/base vocabulary pools, “SUC_basevoc”), it contains valuable information on adjusted frequency (described later), morphological tags for all forms of the lemma, and a running number for frequency range.

Below is an example of the type of information that one can find in the base vocabulary pool.

Table 2. Structure of the base vocabulary pool.

1	2	3	4	5a	5b
38	hon.PF	3261.421389	9	henne.PF@USO@S.817	hon.PF@USS@S.3905

Numbers below correspond to the numbers of Table 2 columns:

- 1 is a running number which identifies this lemma’s frequency range;
- 2 is the lemma, i.e. uninflected form of the word followed by a morphological tag (part of speech). The set of tags is derived from Parole Corpus
- 3 is the adjusted frequency calculated according to the principles explained further in the text
- 4 is the number of text types in the corpus in which this word has occurred (explained further in the text)
- 5 5a, 5b, etc. are different morphological forms of the same lemma followed by their morphological tags and frequencies.

The base vocabulary pool is created on the assumption that domain- or genre-specific words, i.e. those words that occur only in one certain domain or genre, should not be the basis of a base vocabulary pool. The core of such a pool should be constituted by

stylistically neutral general-purpose words collected from as many domains and genres as possible.

Formula for adjusted frequency calculation is given in Forsbom (2006) and takes into account relative size of the genre where the word occurs, its distribution over different language subtypes, frequency of the word in different subtypes and a number of subtypes/genres where the word is used.

Words occurring in less than three genres/domains have been filtered, i.e. no domain-dependent lemmas are used in the adjusted frequency lists (FL). As a result out of 69,371 entries in the total vocabulary based on SUC, only 8,215 lemmas constitute the base vocabulary pool, providing an adjusted frequency list across three category divisions (genre, domain, text). Yet, in spite of a proportionally small number of lemmas constituting the base vocabulary pool, they account for 88.2% of the SUC texts (Forsbom 2006). In the context of second language learning it means that a learner who has acquired the knowledge of these words can read and understand most of the modern Swedish texts.

As a part of this generator, the base vocabulary pool has been split into 8 smaller text files corresponding to 8 frequency bands (FB): 0 - 8000 for easier access to the words of each frequency band.

Following tags are used in the base vocabulary pool for lemmas (more detailed tags, containing morphosyntactic information, are used for different word forms):

Table 3. List of POS tags used in base vocabulary pool

Part of Speech	POS tag
adjective	.AQ
adverb	.RG, .RH
cardinal number	.MC
conjunction	.CC
determiner	.D0, .DF, .DH, .DI
foreign word	.X
infinitive marker	.CI
interjection	.I
noun neuter gender (ett hotell)	.NCN
noun non-neuter gender (en bok)	.NCU
noun shortening	.NC0
ordinal number	.MO
participle	.AF, .AP
particle	.Q
preposition	.S
pronoun	.PF, .PE, .PH, .PI, .PS
proper noun	.NP
punctuation	.F
subjunction	.CS
verb	.V

3.5.1 The FL in yes/no items

In yes/no tests students have to mark individual words (real and nonsense ones) as existent or non-existent in Swedish. The words are presented out of context. For each frequency band (FB) 40 existent words (uninflected forms) are randomly selected from a corresponding FB and 20 nonexistent words (nonsense words) are generated based on the statistics about average number of syllables characteristic for that FB.

Base vocabulary pool contains among other lexical items proper names and numerals that are written with digits. Since for the yes/no test both proper names and words starting with digits (i.e. 1917) are irrelevant they have been filtered from the frequency lists used for this module.

3.5.2 The FL in automatic selection of target vocabulary items from texts

There are several exercise types that can be generated by this exercise generator apart from yes/no items, namely C-tests, wordbank items and multiple-choice items.

C-tests are a type of cloze items where instead of removal of a target word a few initial letters are provided as clues.

In wordbank items all removed items are collected in one table, usually in an alphabetical order. The learner has to match each item with a gap. There are different variations, as for instance to provide more words than there are gaps or to remove only words of the same wordclass so that the learner does not have unnecessary clues. Such variations depend upon the user proficiency level.

Multiple-choice items are items where target words are removed and each gap is provided with several alternatives, among them the right word and a number of distractors.

In C-tests, wordbank items and multiple-choice items the user is provided with several possible options.

1. The first option is to generate items from a manually selected text, with a manual or automatic mark-up of target words.

The use of frequency lists is not needed when the text is manually marked for target vocabulary, see **Figure 3** for the steps taken by the program in generating an exercise:

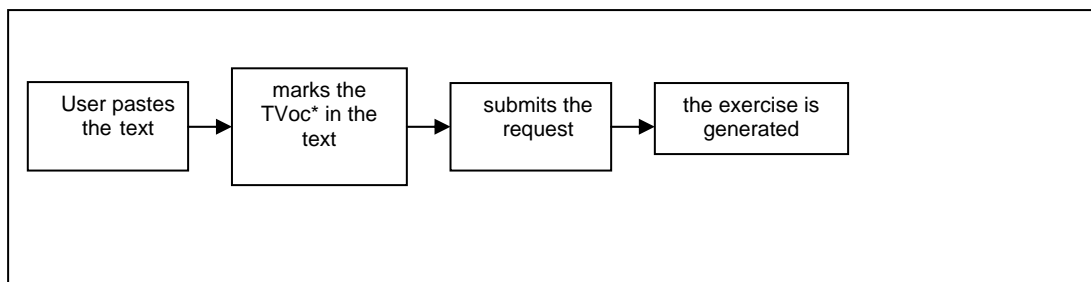



Figure 3. Manual selection of texts with manual mark-up

* Here and in the following figures: TVoc = target vocabulary; TWrds = text words;  marks the use of FL

As soon as automatic mark-up of target vocabulary is required, certain demands are set on the program. Unfortunately, there is no available PoS-tagger or lemmatizer (to my knowledge) for Swedish that can be built into this generator or (re)used as an on-line service. This means that texts either have to be preprocessed before they are given to SCORVEX or to be handled in a naïve way. By the naïve way I mean to neglect the complexity of homonymy across and within the same part of speech and offer the user to disambiguate homonymous words and forms. In an even more naïve approach and possibly even more erroneous one we can use one of the possible parts-of-speech for a certain word without any systematic disambiguation.

In the latter case each text word can be matched against a selected frequency band. Frequency lists contain not only the uninflected form of a word, but even all possible inflected forms that have been used in the corpus. It is therefore easy to find lemma of an inflected form, if that word belongs to one of the Frequency Lists.

Automatic mark-up of target words in texts is therefore FL-based (see Figure 4). Each text contains vocabulary of different frequency bands, but depending upon the difficulty of the text the relative proportion of words from different bands differs. The item-writer has to mark which frequency band he or she wants to train in an exercise, i.e. any of the 4 groupings of frequency bands: 0-1000 words, 1001-2000, 2001-4000, and 4001-8000. Words from a text are matched against the marked FBs; those that match any entry in FL are stored in a separate list and then are selected according to a pattern.

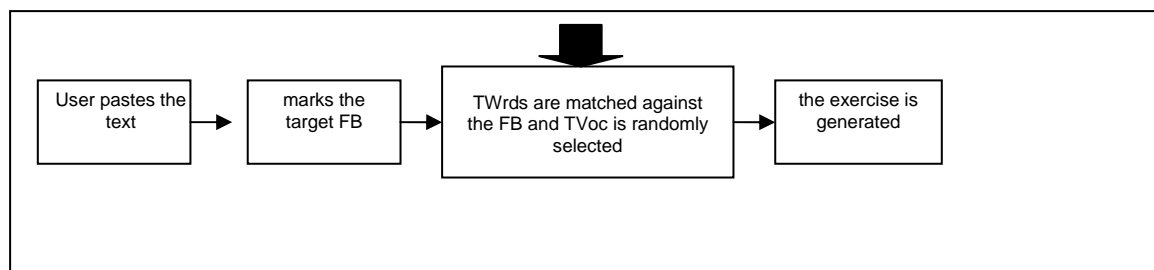


Figure 4. Manual selection of texts with automatic mark-up

2. The second option is to generate items from an automatically selected text with a manual or automatic mark-up of target words

With automatically selected texts the situation is easier. Texts are selected from SUC and are therefore well-annotated. Each lemma is matched against entries in the selected frequency band(s) which guarantees no homonymy across wordclasses; lemmas, textwords and morphosyntactic tags being stored separately until an exercise is created. To match a list of lemmas against a marked FL and then randomly select target vocabulary for training is an easy task then, see **Figure 6**:

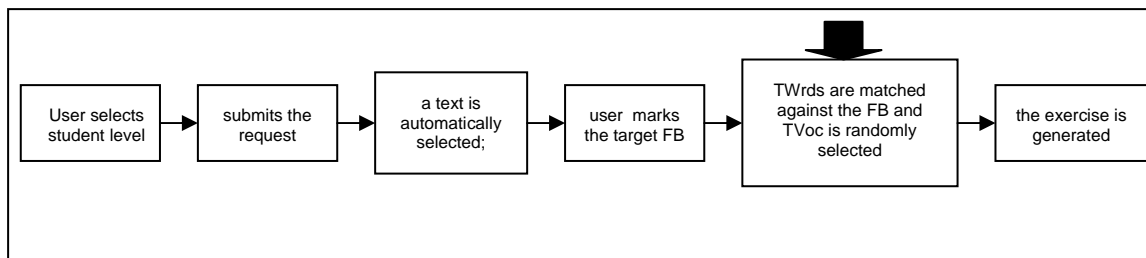


Figure 6. Automatic selection of texts with an automatic mark-up

3. The third option is to generate items from a list of target words.

In this case FLs are not needed for selection of target items (see **Figure 7**). Yet, consulting FLs can become necessary if the items require automatic selection of distractors (e.g. multiple-choice items).

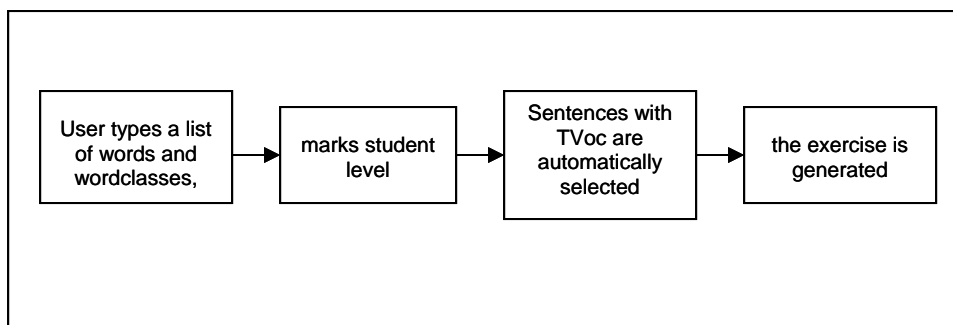


Figure 7. Creating exercises from a list of words

4. One more option is to generate items by random selection of target vocabulary from a specified frequency band (with or without specifying target word classes).

When the test-constructor wants automatically selected words and sentences for the vocabulary item, he or she needs to mark a FB. The user can also specify what parts of speech he or she wants to train (any, only content words, only functional words, or any specific part(s) of speech). Words are randomly selected from the FB (following the restrictions set by the user), their PoS tags are collected, sentences are looked up in SUC, and an exercise is generated (see **Figure 8**).

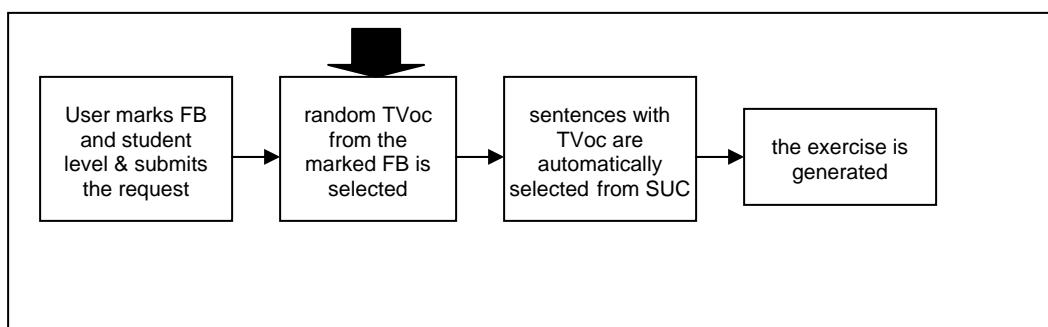


Figure 8. Creating exercises from an automatically selected list of target words

3.5.3 The FL in selection of distractors for multiple-choice items

One of the most interesting uses of FBs and the annotation used in SUC is the process of selecting distractors. As soon as the target vocabulary is identified (manually or automatically), three distractors to each lexical item are selected. First, the target word is checked for its PoS, morphosyntactic tag and frequency band; then the corresponding frequency band is searched for these tags making a list of candidate distractors, finally three words in exactly the same form as the target item are extracted (e.g. for the word “förklaringar” *bakgrunder* och *anledningar* are selected. All the three words share the same tag, namely .NCUPN@IS). **Figure 9** summarizes the steps taken by the program:

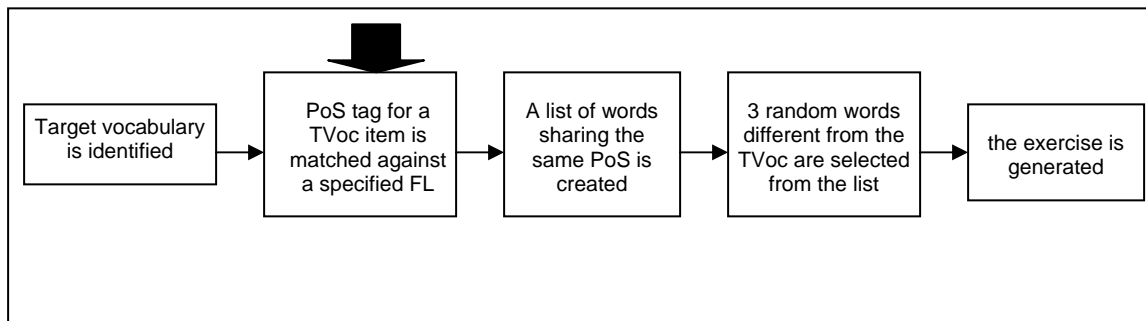


Figure 9. Automatically collecting a list of distractors for multiple-choice items

When target vocabulary is marked manually, the test-producer should set the correct word class to the marked words. The following word classes are accepted by the program:

Table 4. List of POS tags used for manual markup of word lists

Wordclass	Parole tag
adjektiv	.AQ
adverb	.RG
determinerare	.D
konjunktion	.CC
particip	.A
preposition	.S
pronomen	.P
substantiv utrum	.NCU
substantiv neutrum	.NCN
subjunktion	.CS
verb	.V

3.5.4 The FL in search of authentic texts. LFP calculation

The frequency lists have been used in a series of tests on SUC texts for identification of their readability difficulty. Text difficulty index used in this procedure has been based upon LexLIX: readability measure with integrated vocabulary difficulty analysis, i.e. an index where LIX and LFP (lexical frequency profile) are combined. The resulting grouping of texts has been used for text selection.

The frequency lists have been used to collect information on lexical frequency profile (LFP) of each text. Number of words from each band has been summed up and their

percentage calculated. LFP for a text can look like this: 70-9-14-7, which means 70% of all the text words constitute words from FB1; 9% of the words in the text come from frequency band 2; 14% – from bands 3 to 8; and 7% – from band 9. Words that do not match any of the words in the frequency lists are considered specialized words or words of higher difficulty and counted towards the 9th band (difficult words).

For this procedure I have used frequency lists with all proper names and numerals preserved. More on that see under the title “SUC as a source of authentic examples”.

3.6 SUC as a source of authentic examples

Vocabulary is often trained on made-up examples, which has both advantages and disadvantages. Easy to understand structures, clear context and meaning of the target vocabulary can be counted as an advantage at the beginning level. Contrived examples can boast these features. At the same time invented examples are as a rule based upon teachers' intuition and can be misleading. Besides, learners will have to cope with authentic texts sooner or later, texts that no one has simplified to their level. Language examples, be it in textbooks or in test items, should therefore strive to be real-life examples (McEnery & Wilson 2001). Finding authentic examples outside textbooks, however, can be a time-consuming issue, unless the teacher uses a well-annotated corpus.

At the same time it is important that provided (authentic) examples correspond to the level of the learner. Providing authentic but incomprehensible and perplexing examples can scare learners. It is therefore critical to select appropriate examples and texts to stimulate learning instead of stalling it (Fulcher 1997). Thus, the two important characteristics of examples have identified themselves: they should be authentic and they should correspond to the learners' level.

Language-specific general corpora are an irreplaceable resource for collecting authentic candidate texts and example sentences for teaching purposes. However, before using any database (e.g. corpora) as a source of teaching material one should be well acquainted with it. Corpora are an enormous resource of linguistic information, yet there are a number of constraints coming with them. It is not only a question - analysis of what particular kinds of linguistic phenomena the features, like annotation, facilitate. It is also a question about what the topic of a randomly selected text is, what its difficulty level is and a number of other aspects that can be effectively used in NLP tools.

Users must know what's inside the database or corpus if they are to properly interpret the data drawn from it. 'Know thy database' is our late twentieth-century commandment to students. Don't be dazzled by its sheer size, and be sure you critically evaluate its appropriateness for the task in hand. p.176 (Peters 1997)

(Wilson 1997) writes that ideally, texts for a corpus used for CALL should be pre-selected with great discrimination. Texts should be graded according to readability, linguistic features, adequate vocabulary, etc. At the practical level this means that (1) an appropriate source of authentic examples and texts has to be identified and (2) the identified source texts have to be analyzed and graded at least according to readability and lexical difficulty.

Automatic selection of sentences and texts from SUC in this generator is made from SUC text files that GU, Språkdata has a license to use. SUC files come in a number of different versions, one of them in Parole format, i.e. tagged with Parole PoS-tags. The version of Parole-annotated corpus has been chosen for this generator since it complies with the tag set used in the base vocabulary pool (which in its turn is used for statistic information about word frequency). Each text (or a group of shorter texts) is stored in a separate file, the length of each file varies between 2118 tokens and 2817, average being 2390 tokens. Texts are in plain text format containing xml-tags so that automatic analysis of different linguistic and extra-textual features can be possible.

Having identified what corpus to use, it is now necessary to decide on what principles texts for language training purposes can be selected. SUC texts can be of different length, different structural and lexical complexity. They are not annotated for readability index or their appropriateness for language learner levels. Selecting texts at random not checking them for their readability and lexical complexity can result in production of inappropriate teaching materials. It is therefore necessary to find out whether SUC texts can address learners of different levels, i.e. test them for readability and lexical difficulty.

Below I describe the readability measures and lexical difficulty in general, which is followed by a description of the tests that have been run on SUC texts and test results.

3.6.1 Readability Indices

When it comes to deciding on text appropriateness for language learners, even experts cannot agree with each other when grading text difficulty independently of each other (Fulcher 1997). Decisions about text difficulty made by teachers are made on intuitive grounds and are often subjective and inconsistent. Simple measurements used in readability formulas, like sentence and word length, though seemingly unimportant and superficial, have proven to be good predictors of text difficulty. Readability formulas can ensure objectivity and consistence to text difficulty analysis compared with human judgements.

There are different measures that can be used to assess readability of a text. This is usually made by using statistical analyses to investigate a number of textual variables that are claimed to influence readability of a text. Different researchers choose different number of statistic variables influencing text features, combining them either into some formula that helps them calculate a certain index or leaving them separately for comparison. A comprehensive presentation of different approaches to readability analyses is given in (Cedergren 1992). Certain approaches are based on purely statistic calculations; others are dependent on human analysis and interpretation.

The common feature of all readability formulas is the fact that they use statistic information about syntax and lexical complexity, which is then interpreted, results are tested against some reference method, and a number of regression coefficients / constants are selected to gain corresponding values.

In Swedish a readability index called LIX is widely used. LIX stands for “Läsbarhets IndeX” (Eng: readability index). It was first offered by Björnsson in 1968 (Cedergren 1992). It differs from formulas for English in that it uses neither regression model, nor does it contain any coefficients. It is based on the percentage of difficult words, amount of words and sentences in the text, punctuation being excluded from the calculation of words. Difficult words are defined as words containing more than 6 letters. The formula looks as follows:

$$\text{LIX} = \text{number of words}/\text{number of sentences} + (100 * \text{number of difficult words})/\text{number of words}.$$

The index then shows what difficulty level a text has:

20 - 25 Very easy, children books

26 – 30 Simple texts, popular magazines

31 - 40 Normal prose, fiction

41 - 50 Average difficulty, normal newspaper texts, fiction

51 - 60 Difficult texts: formal, expert/factual texts
over 60 Very difficult texts, e.g. research, thesis, bureaucratic language

Another Swedish readability index is Rix which is a simplified form of LIX devised by Anderson (Cedergren 1992):

$Rix = 100 * \text{Number of difficult words} / \text{number of sentences}$

This formula has been used in pedagogy to identify text difficulty level corresponding to school levels.

It has been discussed in literature whether it is appropriate to apply readability formulas used for native speakers on the texts aimed at L2 learners. There have been mixed results, some studies showing that formulas have to be adjusted to a new target group, others providing evidence that the same formulas can be applied to second language learners as well as to native speakers (Greenfield 2004). We assume, following Greenfield's evidence, that the same measures can be applied to L2 learner texts.

3.6.2 Lexical Difficulty Measures

The majority of text researchers work under the assumption that the main factors determining text difficulty are grammar, syntax and vocabulary; syntax and grammar being given the leading role. It is, however, argued by a number of researchers on reading comprehension and vocabulary in L2 that complexity of the vocabulary is the best predictor of text difficulty (Laufer & Nation 1995; Alderson 2000; Read 2000; Meara 2005), far better and more important than grammar and syntax. Difficult vocabulary, even used in short and structurally simple sentences will make it difficult for the learner to understand the text, whereas structurally complex sentences with simple vocabulary will be understood by the learners. In the literature one can see that certain steps have been taken towards combining statistical approach and lexical measures in modeling language that is specific for different learner grades (Brown & Eskenazi 2004; Collins-Thompson & Callan 2004; Heilman et al. 2007).

A number of well-known lexical measures are frequently used in text analysis. Among those measures there are:

- lexical density (LD)
- lexical variation (LV)
- lexical frequency profile (LFP)

Lexical Density (LD) demonstrates a proportion of lexical (content) words in the text. The higher the number, the more written-like the text is. This measure is claimed to be useful when differentiating between spoken and written mode of language. LD is calculated as follows:

$LD (\%) = (\text{total nr of lexical words} * 100) / \text{tokens}$

What should be considered lexical words is continually discussed by researchers. Items like *put up with* are calculated as one lexical unit by some researchers, and as three different items by others. Other peculiarities can appear as well (Read 2000).

tokens in the formula above stand for the total number of running words, punctuation excluded. The same word form that appears more than once is calculated according to the number of its occurrences.

Lexical Variation (LV), is a measure of how varied one's language is, i.e. how many different words a writer uses in his/her writing. This is calculated in the following way:

$$LV (\%) = (\text{types} * 100) / \text{tokens}$$

types are words used in the text counted only once (unique words), excluding punctuation. Uniqueness is secured through lemma-calculation, i.e. even different inflected forms of the same word (the same wordclass, too) are counted as one type, e.g. *pojke, pojkar, pojarna, pojkes*, etc. A high LV number indicates that the writer has a rich vocabulary, whereas the low number shows that the writer relies on a small vocabulary, repeated frequently in the text.

A simplified variant of LV is type-token ratio, where types are unique words, lemma notion not being taken into account.

The drawback of the two above-mentioned measures (LD and LV) lies in the fact that they are highly dependent on the length of the text. The longer the text is, the lower their values get. They are therefore rather unstable measures to be used as predictors of text difficulty.

Lexical Frequency Profile (LFP) is a lexical profile of an individual text, more closely a measure of relative proportion of words from different frequency bands in a text. All lemmas from a text are tested against frequency lists and information is collected as to how many words of each frequency, calculated in percent, have been used in the text. Words that are not in any of the frequency lists are considered to be of lower frequency (rare words) which means more advanced vocabulary. The measure has originally been offered by Nation as a way of assessing suitability of a text for a particular proficiency level of a language learner (Meara 2005). Later, LFP has been taken a step further by (Laufer & Nation 1995) and offered as a measurement for assessing vocabulary used in writing by language learners.

LFP is, thus, a representation of text vocabulary complexity presented in percentage of tokens from different frequency bands. The main distinction is made between words of first 1000 most frequent words, second 1000, the University Word List and any other vocabulary. For more advanced learners finer distinctions may be introduced.

Laufer and Nation discuss advantages and disadvantages of the above mentioned measures (Laufer & Nation 1995). According to them LD and LV, although providing certain measures of lexical variation or richness in the text, do not convey the information what frequency bands the words belong to. The text may contain a lot of different words, all belonging to the same frequency band. Applied to our particular aims, these measures, taken abstractly, do not tell us whether the learner is ready to understand a text as far as its vocabulary is concerned, to say nothing of the appropriateness of using such a text for e.g. generation of gap cloze items. The LFP, on the other hand, delivers the information that seems to be useful in the context of assessing vocabulary suitability for language learners.

Since LFP has been offered as a measurement tool of lexical knowledge, it has been extensively used to create text vocabulary profiles (Meara 2005); LFP has even been tested by different SLA researchers as a measure of assessing progress in learners' vocabulary growth. The most criticized issue is the fine difference between the values that are above the first two FBs. Since most frequent are the words from the first and the second bands, it is not too much space left for distinction between more advanced vocabulary (Meara 2005).

When applying LFP to professionally written texts, with the main aim to evaluate how advanced their vocabulary is, it feels more appropriate to make finer distinction between FBs. Clearly, the majority of text words will be constituted by the words from the first two frequency bands. The difference between less advanced and more advanced texts will assumingly be seen in the tiny difference of percentage of words above the mentioned frequency bands. To make that difference clearer, it might be useful to filter away "stop words" from frequency bands prior to LFP calculation, i.e. discard most frequent words that do not add to the index of lexical complexity. The idea of filtering "noise words" altogether and not count them towards frequency statistics has been also expressed by some corpus linguists (Minugh 1997).

The notion of stop words, otherwise called noise words, comes from information retrieval and stands for the list of words that are filtered prior to processing of search requests. Noise words are believed to make the search query difficult to interpret. Lists of stop words differ from language to language and even within the same language community implementers of different search engines incorporate different lists of stop words. Regularly these consist of most frequent function words. Even some very frequent content words can be added to the stop list.

3.6.3 Test setting

The outset for the test has been defined by the following questions:

1. What should be included into readability difficulty?
2. Which words to consider "noise" words?
3. What to include into LFP and according to what principles to collect the statistics?
4. How to treat "advanced" vocabulary?
5. What place should readability measures vs lexical difficulty measures be given?

1. What should be included into readability difficulty?

Intuitively, text difficulty depends upon complexity of grammatical constructions, syntax, and vocabulary. Almost all known readability formulas take into account sentence length, word length and/or number of syllables. Even though it seems really shallow to predict readability difficulty by calculating long words and sentence length, readability formulas, e.g. LIX, have been used for a long time, and with success, one example of such use is (Nilsson & Borin 2002). Therefore, to avoid reinventing a wheel, I have decided to use LIX as a reference index (and for assigning difficulty levels to SUC texts).

Yet, the question of vocabulary difficulty of a text loomed over me, until I have decided to collect all possible information from SUC and to see how lexical variation, lexical

density, and more importantly, lexical frequency profile correspond to LIX measures and how, if at all, they reflect the difficulty of a text. I have thus decided, apart from LIX, to collect information on LD, LV and LFP measures to get a picture of lexical difficulty.

The most interesting issues in this test have become:

1. Can SUC offer texts for every learner level (from beginner to advanced)?
2. How do text vocabulary profiles (LFP) correlate with text readability indices (LIX) – is there any predictable tendency? If there is, then readability measures do not have to be complemented by vocabulary profiling; if not – there is the next question:
3. What should be done to make automatic text selection more “vocabulary difficulty aware”?

2. What to consider “noise” words?

LD and LV are clear-cut measures, and I have used their formulas without any innovations from my side. LFP, on the other hand, offers some room for experimentation, as can be seen from the description in the section above.

Identifying noise words has turned out to be a tricky task. What can logically constitute noise words in lexical difficulty measures of instructional texts? Which words can impede differentiation between texts with easy and difficult vocabulary?

The first idea was to sort out all functional words, since they do not add much content information to the texts, being useful mostly at syntactic level. At the same time not all functional words should be blindly discarded from lexical difficulty analyses, e.g. words like “alltmedan” (subjunction) are rather advanced and relatively rare.

The first “noise candidates” have become functional words from FB1. I have run a on frequency lists to count how many functional words each FB contains. Filtering of functional words has been done automatically by using PoS tags. The following word classes have been counted towards function classes (and thus “noise”) (see Table 5):

Table 5. List of functional wordclasses⁶

Wordclass	Parole tag
conjunction	CC
determiner	D0, DF, DH, DI
infinitive marker	CI
punctuation mark	FE, FI, FP
preposition	S
pronoun	PF, PE, PH, PI, PS
subjunction	CS

⁶ After the calculations have been made my attention has been directed to the fact that shorter adverbs of the type “dit”, “här”, “ut”, “nu” etc. should also have been included into the calculations of functional words. Even interjections and particles should have been counted among functional wordclasses.

Counting the amount of function words vs content words per band in the eight frequency bands has yielded the following results:

Table 6. Number of functional words per frequency band

Frequency Band	Number of functional words	Relative proportion (in %) per 1000- band
1	105	10,5
2	30	3,0
3	22	2,2
4	17	1,7
5	14	1,4
6	7	0,7
7	4	0,4
8	4	0,4
Total:	203	2,47 % (per 8 bands)

Appendix 2 contains all functional words sorted by frequency band.

All in all, there are 203 different function words spread over eight frequency bands, whose total number of lemmas is 8215. As it can be seen from Table 6, number of function words per frequency band declines drastically after FB1, which is a predictable tendency. However, the absolute number of function words is too small to be able to solve the problem of discriminating between words from more advanced frequency bands by filtering them from calculations.

The next idea that looked reasonable was to calculate relative proportion of words from band 1 as opposed to all other words, and afterwards zoom in all other bands except FB1 and calculate their relative proportion. Bands 2-8 become thus a focus of more close examination, whereas FB1 is given the status of a “noise band”. Presumably, even beginners, when they start reading texts, will already have the knowledge of the first 1000 most frequent words in Swedish. The calculations have shown that words from FB1 constitute approximately 69,9% of the whole SUC corpus, the value span for each text differing between 48,4% and 83% per text. By distracting FB1 words from calculation of LFP, it seemed that we can see more clearly the difference in vocabulary frequency profile between more advanced texts and less advanced ones.

Yet, this approach has also shown to be fruitless. The most important reason is that the numbers received from “the rest of the vocabulary” in a text (i.e. relative proportion of words from FBs 2-9) cannot be compared from text to text since FB1 takes up alternatively different percent of the text.

In the end straightforward numbers from each band have been collected with the intention to analyze them and afterwards draw conclusions.

3. What to include into LFP and how to group FBs

The fact that FB1 constitutes the major part of the lexical profile in any text makes it very distinct. Therefore there has never been a question whether words from FB1 should be grouped in calculation with any other FBs. However, the remaining words in the text can either be collected according to each frequency band from FB2 to FB8, or FBs can be grouped in some way. Obtaining 7 values for FBs 2-8 does not seem reasonable; too

many numbers can be confusing to deal with and interpret. I have therefore tested clustering FBs in the following way: frequency band 2, 3-4, 5-6 and 7-8. After the data have been collected, I clustered bands as follows: FB1, FB2, FB3-8, FB9+. The division is arbitrary and may need to be adjusted in some way. Regrouping of already available numbers is also possible with this approach.

4. How to treat “advanced” vocabulary?

Another practical “technical” problem has been: what to do with the words that do not belong to any of the eight frequency bands? Nation (Meara 2005) collects all such words into “other words”. I have decided to call them words from frequency band 9-plus (9+), or “advanced” vocabulary. What words should be counted towards “advanced” vocabulary? Among those words that do not belong to any of the eight FBs there are numerous proper nouns and (ordinal) numerals with digits that are hardly “advanced words” even though they do not belong to any of the eight frequency bands. Should they be counted as “more advanced vocabulary” or filtered away? I have so far decided not to count proper names and numerals towards band 9+, though they are counted towards running words.

5. What place should readability measures vs lexical difficulty measures be given?

This particular question is the topic of the next section.

3.6.4 Test results, generalizations and conclusions

The following LIX numbers have been obtained for SUC texts:

Level 1 (LIX value up to 25)	34 texts;
Level 2 (LIX value 26-30)	39 texts;
Level 3 (LIX value 31-40)	167 texts;
Level 4 (LIX value 41-50)	181 texts;
Level 5 (LIX value 51-60)	74 texts;
Level 6 (LIX value 61+)	5 texts;
Total	500 text files

The result is very encouraging; it tells us that appropriate texts can be selected as good as for any proficiency level from SUC. The majority of SUC texts correspond to LIX levels 3 and 4 (fiction, normal newspaper texts), most underrepresented are highly difficult texts corresponding to LIX level 6 (only 1% of all texts). Easier texts (LIX levels 1 and 2) constitute 6.8% and 7.8% respectively of all SUC texts. Now that each text file has a readability index, it is an easy task to automatically select texts of a necessary difficulty level.

The correspondence between LIX values and LFP is not direct. Numbers received for bands 3-8 are so low that I have decided to group them together to achieve better representativity. Table 7 contains average values per LIX level and band:

Table 7. Average values per LIX level and frequency band

LIX levels	Band 1	Band 2	Bands 3-8	Band 9+	LD	LV
Level 1	76.5 %	6.8 %	8.2 %	8.5 %	54 %	32 %

Level 2	74.8 %	6.8 %	9 %	9 %	54 %	34 %
Level 3	71.2 %	7.4 %	11 %	10.4 %	57 %	37 %
Level 4	68 %	8 %	12.5 %	11.4 %	58 %	37 %
Level 5	66.7 %	7.9 %	12 %	13.3 %	59 %	35 %
Level 6	61.5 %	7.7 %	12 %	18.6 %	59 %	38 %

The distribution of words from different vocabulary bands per each LIX level, in average, is given even in appendix 3 in a pie diagram form.

It is obvious that as the readability index grows (which means the texts become more difficult), the relative proportion of FB1 words diminishes whereas words from band 9+, on the other hand, show a stable tendency to increase as LIX grows. Words from FB2 remain at approximately the same level. Words from FBs 3-8 increase from LIX level 1 to 4, and remain at the same level at levels 5 and 6. On level 6 the amount of rare words from band 9+ is at its maximum (18,6 percent). Lexical Density and Lexical Variation grow, too (not significantly, though).

The values seen in the table above are mean values per LIX level. The value span for FBs in each LIX level is very big and values are overlapping between LIX levels, as can be seen from Table 8:

Table 8. Value span for each FB and LIX level

LIX levels	Band 1	Band 2	Bands 3-8	Band 9	LD	LV
Level 1	63.2 - 83	4.6 - 9	5.8 - 11	4.4 - 18.7	47 - 63	25 - 39
Level 2	57.4 - 81.6	4.8 - 10	6.3 - 12.6	5.13 - 20	46 - 66	29 - 41
Level 3	50 - 82	4.7 - 11.6	6 - 18.5	3.4 - 27.5	48 - 65	19 - 47
Level 4	55.5 - 79.6	4.2 - 13.4	6.7 - 20	5.3 - 25.2	51 - 63	17 - 52
Level 5	48.4 - 77.2	3 - 14	6.2 - 26.5	6 - 26.3	44 - 65	24 - 51
Level 6	56.4 - 67	6.7 - 9	10.3 - 15.8	13.3 - 10.3	54 - 64	35 - 42

Analysis of those values allows making conclusions that proportion of words from bands 1 and 9+ and their relationship are the most reliable factors in predicting lexical difficulty of a text. The higher the proportion of FB1 words is, the easier the text is, and vice versa, the fewer words from FB1 compensated by a high proportion of words from FB9+, the more difficult the text is, lexically viewed.

Subtracting percentage of FB 9+ words from FB1 words and then from 100 gives us a number that presumably can function as an LFP score ($LFP \text{ score } (\%) = 100 - (FB1 (\%) - FB9+ (\%))$). The lower the LFP score is, the easier vocabulary the text contains (which means numerous words from FB1 and relatively few words from FB 9-plus). The same tendency show even LIX values. Low LIX values, as well as low LFP-score values point out easy texts.

The span of values for LFP-scores for SUC texts varies from 23 to 79. The lexical profile (LFP as it is) can look like this: 82-7-8-3, which means word distribution according to FB1-FB2-FBs3/8-FB9+ shown in percent.

I have gone further and combined the two indices (LIX and LFP-score) into a new “vocabulary aware” readability index that I have called *LexLIX* (lexical LIX):

$$\text{LexLIX} = (\text{LFP-score} + \text{LIX}) / 2$$

Summing them up and then dividing into 2 helps introduce corrections for lexical difficulty of each text into LIX value. The same grouping of scores into difficulty levels as used for LIX can be (presumably) applied to LexLIX. At least that is what I have decided to test.

The detailed analysis of correlation between LFP-score, LIX and LexLIX values is not the topic of this thesis, but I can name some interesting facts I have discovered during the first examination (skimming) of the numbers.

In Table 9-11 the ranking of easiest versus most difficult texts is provided according to the three indices:

Table 9. Easiest and most difficult texts ordered by LIX

	LIX	LFP-score	LexLIX
Easiest			
kk09	18	26.3	22.2
kk10	21	29.7	22.7
kk59	21	29.5	25.2
kl15	22	36.1	24.7
Most difficult			
jc19	62	54.3	59.7
ja04	62	53.3	63.2
jc05	66	61.7	63.8
ja14	67	70.0	68.5

Table 10. Easiest and most difficult texts ordered by LFP-score

	LIX	LFP-score	LexLIX
Easiest			
fa02	35	21.2	28.1
kk70	24	21.4	22.7
fb01	33	21.9	27.5
fb02	33	22.3	27.6
Most difficult			
ja08	43	66.0	54.5
jg02	58	68.4	63.2
ja14	67	70.0	68.5
ha23	38	77.4	57.7

Table 11. Easiest and most difficult texts ordered by LexLIX

	LIX	LFP-score	LexLIX
Easiest			
kk09	18	22.3	22.2
kk70	24	21.4	22.7
kk13	23	22.5	24.7
kl09	22	27.5	24.7
Most difficult			
ja05	54	65.4	59.7
kg02	58	68.4	63.2
jc05	66	61.7	63.8
ja14	67	70.0	68.5

The text that is lexically the easiest according to LFP-score as can be seen from table 10 (text fa02) has LIX 35, i.e. level 3, normal text difficulty. It is a text about communication and personal identity – an easy-to read and understand text that reminds of an argumentative essay. The text that has been identified as the easiest according to LIX (kk09) as shown in table 9 has LFP-score 26,3. The second of the two texts (kk09) is in my subjective opinion easier, it is an extract from imaginative prose and has a lot of dialogues and inner dialogues, and is full of spoken language. The fact that LFP-score has not identified this text as the easiest one is based presumably upon the sad truth that frequency bands have been derived upon written language. Expressions typical for spoken language like “fan” (Eng. damn), “tapeter” (eng. wallpaper), “kängor” (Eng. boots), “kackla” (Eng. to cackle), “full” (Eng. drunk), etc. are not among the words of the first 8 frequency bands. These easy words have obviously been calculated towards the 9-plus band in LFP-score. To avoid such distortion it would have been useful to engage some list over frequent spoken words and expressions and test each potential 9-plus entity against this list before allowing it to be calculated towards 9-plus band.

Combining LIX and LFP-score into a new index helps to bring kk09-text to the first place, i.e. give it a status of the easiest text. LFP-score is thus corrected for structure and syntax, whereas LIX is corrected for lexical difficulty.

The matter is a bit different when it comes to the analysis of the most difficult texts according to LIX and LFP. The text, identified as the most difficult by LIX (ja14), has the corresponding LFP-score 70 which points at an extremely difficult text as well. The text abounds in foreign words and special terminology; and obviously they account for the high LFP-score for the text. The text identified as the most difficult by LFP-score (ha23), on the other hand, has a corresponding LIX value 38 – normal prose, easier texts. It has shown, however, to be a law text consisting of terminology and is not appropriate for language learning purposes whatsoever. It is most probably a matter of suitability rather than readability.

Combining both indices brings ja14 to the “most difficult text” place, and ha23 gets the LexLIX 58, and is the 7th most difficult text in SUC collection.

I have decided to see how the ranking of texts by LIX vs LFP corresponds to the ranking made by human readers. To do this I have asked 7 persons to read a selection of 9 SUC texts and asked them to order them from easiest to the most difficult using their intuition and paying attention to vocabulary, grammar and syntax. The extracts from the texts are

provided in appendix 4. The results have confirmed what has been said in the beginning of the chapter on readability indices: human readers do not agree with each other when grading texts for difficulty separately from each other. Tendency, though, is common for both automatic algorithms and human intuition: easier texts are placed at the beginning (in different order, yet all the four ones are on the top in the majority of cases), difficult texts are placed at the end.

Table 12. Ranking of texts graded for difficulty by human readers from easiest to difficult

Lix	LFPsc	LexLIX	R1 (n-N)*	R2 (n-N)	R3 (n-N)	R4 (n-N)	R5 (N)*	R6 (N)	R7 (N)
kk59	fa02	kk13	kk59	kk59	bb01	ha23	kk59	kk52	kk59
kk13	kk13	kk59	kk13	kk13	kk59	kk59	kk13	kk13	kk13
kk52	kk59	fa02	kl19	kk52	kk52	kk13	kl19	kl19	kl19
fa02	bb01	kk52	kk52	kl19	ha23	bb05	kk52	bb05	kk52
kl19	gb17	kl19	gb17	fa02	bb05	fa02	fa02	kk59	fa02
ha23	kk52	gb17	fa02	bb05	kk13	kl19	bb05	bb01	bb05
gb17	bb05	bb01	bb05	bb01	fa02	bb01	ha23	fa02	gb17
bb01	kl19	bb05	bb01	gb17	gb17	gb17	gb17	gb17	bb01
bb05	ha23	ha23	ha23	ha23	kl19	kk52	bb01	ha23	ha23

* R(1-4) = Reader; n-N = non-Native speaker of Swedish; N = native speaker of Swedish

Table 13. LIX, LexLIX and LFP-scores in the 9 human-graded texts

Text	LIX	LFP-score	LexLIX
kk59	21	29	25
kk13	23	26	24
kk52	29	40	34
fa02	35	21	28
kl19	37	40	38
ha23	38	77	58
gb17	41	40	40
bb01	44	38	41
bb05	45	40	42

As can be seen from Table 12, the LIX and LexLIX scores are distributed in such a way that there are clear-cut groups of texts that have similar (or close) difficulty scores. Easiest are kk59 and kk13 with very slight difference in scores; next come a group of kk52, fa02 and kl19; the third group is constituted of gb17, bb01 and bb05; in its own class is ha23 since the two indices have given it different scores. Within each class human readers have made some modifications in order, yet the order of difficulty classes is preserved in almost all the rankings. But clearly, most human readers have agreed with the LexLIX estimation that ha23 is the most difficult text.

The ordering can also be explained by the genre of each text. Those that start with letter “k” belong to imaginative prose and are therefore easier to read (reader-friendlier), whether the syntax and vocabulary are slightly more difficult or not. It is usually typical of fiction to have dialogues which consist of short sentences which directly influence LIX

score. Texts that start with “b” are examples of editorials and are therefore less entertaining in their nature, contain more specific vocabulary that are not so colloquial in nature and have in general longer sentences – aspects that influence both LIX and LFP-scores.

I view these results as very encouraging. Obviously, combining LIX and LFP helps bring together two important measures: syntax on the one hand and vocabulary difficulty level on the other hand. LIX and LFP-scores compensate each other and introduce corrections into each other’s scores. I would like to test using LexLIX score as a primary readability index for automatic text selection in the exercise generator. Grouping into levels according to LexLIX has been done in the same fashion as for LIX:

-25 very easy texts; 26-30 - ...etc. Then we can obtain the following numbers:

Level 1 (LexLIX value up to 25)	12 texts;
Level 2 (LexLIX value 26-30)	45 texts;
Level 3 (LexLIX value 31-40)	189 texts;
Level 4 (LexLIX value 41-50)	203 texts;
Level 5 (LexLIX value 51-60)	48 texts;
Level 6 (LexLIX value 61+)	3 texts;
Total	500 text files

The various statistics for SUC texts has been collected and saved in an Excel file. For those interested in it, mail the author at <elenavolodina@yahoo.com>.

Conclusions

The first question that this study strove to answer - if SUC is an appropriate source of learner texts for different levels – can now be answered. The answer is – definitely yes. It is possible to select texts of different readability levels and different genres.

The second question about the correlation of lexical measures and LIX can be answered as well. There is no linear dependency; though the scores received by LFP and LIX point at approximately the same difficulty level, yet not as straightforwardly as I believed they would.

The third question to be answered is how can we make automatic text selection more “vocabulary aware”? The first step towards that answer has already been offered. LexLIX seems to be a good alternative to LIX, though a series of serious tests need to be run to test this measure further. Intuitively, however, I believe that LIX with corrected score for vocabulary difficulty is an appropriate index for selecting texts for L2 learners of Swedish.

3.6.5 Algorithm for text selection.

Now that the grouping of SUC texts is made into levels, automatic selection of texts is done according to the following scheme (algorithm):

First, one text of the appropriate level is randomly selected from a list of texts belonging to that level;

Second, since each text is about 2300 words long, only a part of the text need to be selected, and that extract has to be coherent. To ensure that the text is at least in some way connected, only full paragraphs are extracted. The excerpt starts at a randomly selected paragraph with every word counted. An extract from between 150 and 250 words is appropriate for any exercise. Therefore as soon as the word count passes the count of 150 the program looks for the end of a paragraph. In the cases when a SUC file consists of a number of shorter texts, the number of texts is calculated, and one of them is randomly selected. Then the procedure above is repeated.

By selecting texts according to the procedure described above we make it possible to automatically select texts that match the student competence level in language skill (as defined by a teacher).

3.6.6 Algorithm for sentence selection

Obviously, it is not only texts that need to be tested for complexity. Sentences as well have to be analyzed for structural and lexical complexity. In certain situations (at a beginner level) long and complex sentences may be unacceptable for language learning purposes since they may inhibit understanding. And vice versa at a more advanced level they may be useful for training in spite of their difficulty.

In the cases when a sentence with target word is looked up in SUC, a specifically designed archive for such search has been created. All SUC texts have been automatically analyzed and an index over all sentences consisting of files named after a lemma plus part-of-speech tag has been created. In each file sentence id-numbers (which include even text/file names and running sentence numbers within each text) are listed followed by text level (LexLIX level). E.g. in the file with the name `folkskola.NCU.txt` there is the following list of sentence ids:

Figure 10. SUC-sentence index. Content of the file “`folkskola.NCU.txt`”

```
<s id=ga07-046 level=3>
<s id=ed01a-012 level=3>
<s id=ab03c-015 level=3>
<s id=ad04a-024 level=4>
<s id=ec10b-039 level=4>
<s id=cc03e-007 level=4>
<s id=jc04-003 level=4>
<s id=jc04-009 level=4>
<s id=jc04-066 level=4>
<s id=jc04-073 level=4>
<s id=jc04-081 level=4>
<s id=jc04-092 level=4>
<s id=jc04-112 level=4>
<s id=jb06-148 level=4>
<s id=jc03-055 level=5>
<s id=jd01-037 level=5>
```

Each id is constituted of a file name and a running sentence number in the text. Extracting the filename from the list of sentence-ids we make it sure to find a sentence with the target word.

When a sentence-id is automatically selected, a corresponding file is opened (e.g. "folkskola.NCU.txt"), from a list of available sentences only sentences of the desired level are selected, and one of them is randomly chosen for an exercise. In the exercise the target word can be used in any form, inflected or uninflected.

It is probably worth mentioning that the archive of sentence-ids contains 69,200 files, which corresponds to the amount of lemmas in SUC (69,371) minus a number of lemmas that start with citation marks and punctuation marks.

4. Vocabulary Generator – Pedagogical Prerequisites, Theoretical Questions and Design

Below I am going to describe the types of vocabulary exercises that can be automatically generated by this system and algorithms for their generation (i.e. contents) neglecting the way they can be presented (i.e. form or format). The issue of adaptivity of a system to a student proficiency level as well as modeling student's competence in vocabulary has been left for future work.

Thus, the objectives of this chapter are to describe theoretical and practical issues of each vocabulary item type including arguable points that had to be solved before the implementation could be carried out. The implementation issues proper are provided in appendix 6.

This computer-assisted exercise generator can at present produce the following items:

- c-test items
- multiple choice items
- wordbank items
- yes/no items

4.1 General information on the gap cloze test items

Cloze procedures are exercises or tests where words are systematically deleted from the sentences, leaving learners with the task of filling in an appropriate word into the gap. This type of test item has been widely used since 1970s in an attempt to step away from decontextualised vocabulary test items common before (Read 2000). There is still a lot of research into its nature and the nature of what it measures. The common assumption is that cloze procedures cannot be viewed purely as a lexical measure. However, learners have to use vocabulary knowledge to a stronger degree than other areas of language proficiency to be able to fill in the gaps.

The original cloze test consisted of a number of reading passages with words deleted according to some specified pattern (e.g. every seventh word). It was used to test readability of texts for L1 students (Read 2000). Later the test attracted the attention of L2 teachers and researchers and gradually came into use in this area.

There exist several variants of the cloze test:

- classical cloze/rational cloze (with words deleted according to some pattern, i.e. every *n*th word)
- C-test
- multiple-choice cloze
- wordbank items

There is an issue of context dependence that arises in connection with cloze items. Is one sentence a wide enough context to guess the word or should a longer context of a passage be drawn? Many researchers point out that some of the blanks can only be successfully filled if the learner is provided with wide context.

There are several possibilities to provide other than contextual clues to the learner. One is to leave one or several initial letters of the word, or even half the word (C-test); another one is to have multiple choice options (multiple-choice cloze); the third one is to group all deleted words in a list (wordbank) offering the learner to choose an appropriate one for each gap. Having no clues at all tests, according to some researchers, knowledge of syntax in greater degree than vocabulary (see discussion around it in Read 2000, p.105). A variant of a classical cloze (where words are deleted according to a pattern) is rational cloze, i.e. deleting words selectively, choosing only content words which can be reconstructed on context clues (rather than on syntactic clues). It is claimed that both classical and rational cloze tests can be used for testing different aspects of L2 proficiency: lexical, grammatical/syntactic, extra-textual, reading comprehension.

Debates into what aspect of language proficiency cloze tests measure continue even today. Some researchers claim that it is only the vocabulary knowledge and “local” grammar that is tested, others argue that it can be used as a learner’s overall proficiency measure of reading and target language in general (Read 2000). C-tests have shown to correlate well with other vocabulary tests rather than with tests of reading or writing proficiency and thus are considered to be a valid test of vocabulary and grammatical elements. It is also assumed that C-tests are a more appropriate measure of language proficiency for higher-level students. (Chapelle 1994) provided a number of arguments that the best way of using cloze tests as a measure of L2 learners’ vocabulary knowledge is to assume rational approach (select items for deletion rather than delete every *n*th item), choosing only content words, mutilating them (shortening) in such a way that they can be restored using contextual clues.

To sum it up, cloze procedures test not only the knowledge of deleted words, but also the knowledge of content words and syntactic structures surrounding the deleted items and even paragraph organization, spelling (in C-tests), word morphology, etc. In this respect these tests are more embedded than the majority of other test items used for vocabulary assessment.

In our program we have chosen to implement three types of cloze procedures: C-test, multiple-choice cloze and wordbank format of gap cloze tests.

4.2 Computer-assisted generation of C-tests

C-test that is implemented in this software is based on texts, either chosen by the user and pasted into the interface window or automatically selected from SUC (Stockholm Umeå Corpus). There are two options as far as target word selection is concerned: either the test-constructor marks the words him- or herself, or the program selects those words automatically from the marked frequency band. Below follows a description of the automated processes.

4.2.1 Automatic selection of target words

Using Read's (2000) terminology, a selective-deletion model is used in this test. The total amount of words in the text is calculated (N), the number of gaps (G) being equal:

$$G = N / 12;$$

It is a rather arbitrary number, which can be adjusted if necessary. The test-constructor can choose from which frequency band he/she wants to test learners' word knowledge and the program automatically searches for words from those bands in the text. To avoid having gaps following close to each other, an algorithm is used that specifies that gaps can be placed at a distance of minimum 5 words from each other. The program also checks that the same wordform is not deleted twice.

The base-pool vocabulary list is organized into frequency bands, against which text words are checked for frequency information, is organized by lemmas. It contains, however, extra information on different word forms of the same lemma, which makes it possible to automatically check words from the text for frequency information without prior text lemmatization.

The selected words are shortened according to the following principles:

- if the word starts with a consonant, the consonant cluster plus the following vowel or a combination of vowels are printed as a clue, the rest of the word is deleted;
- if the word starts with a vowel, the vowel plus the following consonant cluster are printed as a clue, the rest of the word is deleted;
- if the abbreviation acc. to the rules above is longer than half the word length, the word is cut in the middle;
- if the word is maximum four letters long, only the first letter is provided as a clue.

When the selection is made, the information about whether the word is functional or lexical can be activated; even a desired wordclass(es) can be selected. These constraints are possible to introduce due to the tag information contained in the frequency lists and in the SUC annotation.

4.2.2 Automatic text and sentence selection

In its present form the authoring tool offers a teacher or test designer to select texts of four difficulty levels – beginner, intermediate, upper-intermediate and advanced. As has

been described in the previous chapter it is possible to select texts of different difficulty, starting with beginner level and all the way up. For selection procedure text readability index and lexical difficulty estimation have been combined. Automatic text and sentence selection is a procedure used in all modules of SCORVEX save yes/no test. The algorithm is described in the previous chapter.

4.2.3 Correction for grammar and spelling

The question of scoring method in gap cloze items has been discussed separately for native and non-native speakers. Native speakers are assumed to be able to restore an exact variant of the deleted items in terms of both the choice of word, as well as its spelling and grammar form, whereas non-native speakers might be allowed to make some mistakes (Oller 1973). The latter scoring technique allows varying degrees of correctness, which are rather subjectively determined. This in its turn might indicate that a human grader – native speaker preferably - should do the scoring.

(Laufer & Nation 1995) claim that a word is not in the testee's productive lexicon if he or she cannot use it correctly. A wrong derivative of a word and wrong spelling are not considered "incorrect use".

Using Laufer and Nation's approach, i.e. disregarding spelling, word-building, inflectional and structural use, leaves us with semantic aspects of the word. Is it the only aspect of the word that we assume should be known by a student? What principles should the scoring be based upon – perfect knowledge of the word (correct form, use and meaning) or partial (only one of the aspects is correct)? Should that depend upon learners' level or should it be applied systematically to all levels?

Provided that students have the first several letters as a prompt, all they have to do is fill in the rest of the word. That calls for productive application of vocabulary knowledge: the learner has to use semantic, collocational and grammatical constraints that are imposed on the shortened word by its environment/context, as well as demonstrate the knowledge of its spelling, inflections, affixation, etc. In case the testee cannot spell the word correctly, some measures to guess whether he/she has meant the right word have to be taken, which calls for spell-checking mechanisms.

Different approaches and techniques within automated correction of word spelling are described in (Kukich 1992). Among the general approaches she describes nonword error detection, isolated-word error correction and context-dependent word correction); an isolated-word correction procedure (or, rather, word recognition) is the most suitable for the purposes of this exercise generator. This is motivated by the fact that we have one word typed by the learner and have to compare it to the correct word offered by the program.

Kukich (1992) points out four possible mutations within the word that may happen when the word is known by the user, but is accidentally misspelled: deletion, insertion, substitution and transposition. Another reason for incorrectly spelled word is phonetic. In this case the learner knows the word pronunciation but fails to find correct correspondence between phoneme and grapheme (letter combination). The third reason for incorrect usage of a word is its wrong grammatical form, which occurs in case the

learner knows the word semantics but lacks the knowledge of its grammar constraints in the given context. Finally, the learner may fail in finding the correct word at all.

Thus, the task that the implemented program SCORVEX has to solve is to identify whether the learner does not know the word at all or has made one of the first three mistakes: accidental misspelling, phonetic or grammar mistake. In neither of these cases we can assume that the learner does not have the knowledge of the word whatsoever and thus withdraw the full point for the word. The score should be reduced by some value.

To handle grammar problem, it is enough to collect all possible morphological forms from the base form frequency list used in the program. In case there is still no match, mutations within the word may be checked.

To proceed with automatic recognition procedure it is vital to decide how long we should go in correcting process. Usually spell checkers generate words that lie at a distance of one from the original word (Domeij, Hollman & Kann 1994), which means only one of the possible corrections is introduced. One can, of course, continue guessing increasing the distance, but it is an expensive process in terms of time and efforts. Let's say we stop at a distance of one. Then, the following algorithm is possible.

First, it is important to identify which part of the word that differs. It can be done by comparing the two words character by character from the beginning of the word till the first difference occurs. The position for different characters has to be stored. Then the words can be compared from the end and likewise the index for different characters should be stored. Comparison of the lengths of the substrings between the two indices can give us an indication of whether we have a misspelled word or incorrect word choice.

The seemingly easiest way to spot deletion or insertion is to use the method known as "the longest common subsequence". If the lengths of the two substrings are equal, we can ignore deletion and insertion procedures and test for substitution or transposition. In case of transposition $n-1$ variants will be generated (swapping neighboring letters). In case of letter substitution, n letters in the word have to be tested for $29-1$ alternative letters of the Swedish alphabet. This demands $28*(n-1)$ generated variants.

If the lengths of the substrings are several letters different it can point out a phonetic mistake. In this case it is necessary to have a correspondence table between the original word's graphemes and its phonemes, and then backwards a correspondence table between phonemes and graphemes.

The algorithm offered here is rule-based. Its disadvantage lies in the fact that if we save the generated test/exercise in QTI format, we have to store even all generated correct answers with the score points in a ready-to-use format.

Instead of implementing a spell-check, there is an option to reuse the existing spell-checker STAVA, which is planned to be done in the future.

4.2.4 Calculation of the score

Once we have access to phoneme-to-grapheme and vice versa correspondence table, the algorithm above can be implemented. We assume here an approach that only one type of

mistake is allowed, i.e. either grammar or spelling, in other words we follow the principle of distance one from the correct word.

The scoring procedure, then, can look as follows:

Correct word guessing gives 1 point;

Incorrect grammar gives 0,5 points;

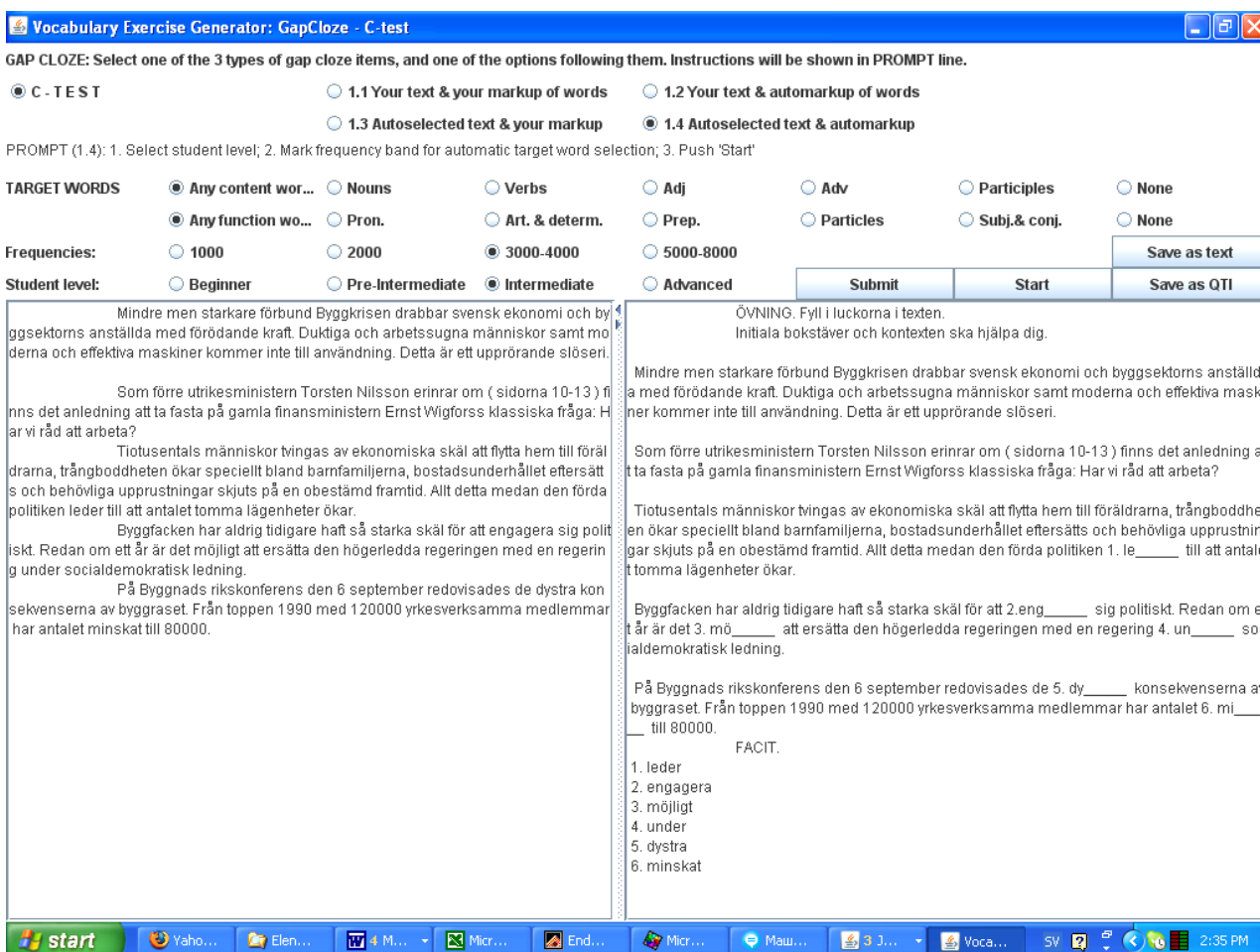
Incorrect spelling gives 0,5 points;

Thus, if a student makes both grammar and spelling mistakes, zero points will be given. This way we underline that the semantic aspects of vocabulary (knowledge of what word to use in a gap) are equally important for the L2 learners as grammar and spelling.

Correction for spelling and grammar, however, has not been implemented into this generator due to the lack of time. It is left as future work. Instead, only 100% correct words are counted as 1 point.

The user interface of the authoring tool looks as follows:

Figure 11. C-test Module, user interface of the authoring tool.



4.2.4 Examples of automatically generated c-items

Merits of the corpus as a source of CALL exercises can be best demonstrated via examples. Therefore in the end of each subchapter on a specific exercise type (except yes/no test) I am showing some exercises that have been automatically generated. These examples demonstrate the span of language learning materials that can be produced on the basis of SUC and this generator.

C-test items test reading comprehension and acquaintance with typical context for target vocabulary items rather than vocabulary as a construct. It is true, that these items are more difficult than the ones described below, since the student has to use vocabulary actively (as contrasted to passive recognition).

C-test module provides several alternatives, all of them being text-based, since large context is of critical importance for c-tests. The words for training can be

- marked manually
- automatically selected from a particular frequency band
- automatically selected from a particular wordclass(es)

Some examples are shown below:

Example 1. C-test: automatically selected nouns for training in a text of intermediate level.

ÖVNING. Fyll i luckorna i texten. Initiala bokstäver och kontexten ska hjälpa dig.

41 Tidigt på morgonen går jag upp på taket. Det är en strålande 1. d _____. De vita molnen ligger samlade vid horisonten som ännu ett sedimentärt lager över svarta berg.

Här stod Fromentin för 130 år sedan. Det är samma 2. la _____ vi ser. Samma sol, samma öken.

Men inte samma människor. Hans araber var slutna, hotfulla, fientliga. De jag har träffat är öppna, levande, gästfria människor.

Under samma 3. s _____.

Fromentin hade fel när han trodde att det var solens obarmhärtighet som för alltid hade präglat öknens 4. mä _____. Kanske visste han det rentav själv. På 1980-talet har En sommar i Sahara kommit i nya vetenskapliga utgåvor som också redovisar textvarianterna. Här finner man andra förklaringar till 5. ty _____ i Laghouat.

42 Våren 1830 kokade Paris redan av det uppror som skulle få sitt 6. utl _____ i julirevolutionen. Den reaktionära kampregeringen de Polignac var fallfärdig. Som en sista utväg för att avleda missnöjet beslöt man angripa Alger. Förevändningen var en påstådd 7. fö _____ mot den franske konsuln.

FACIT.

1. dag
2. landskap
3. sol
4. människor
5. tystnaden
6. utlopp
7. förolämpning

Example 2. C-test: Automatically selected words from FB 3000-4000 in a text of pre-intermediate level.

ÖVNING. Fyll i luckorna i texten. Initiala bokstäver och kontexten ska hjälpa dig.

- Ge nycklarna till Li, mimade Nilla och Katty plockade fram nyckelknippan.

Dörren såg 1.osk_____ stängd ut. Den bar inga spår av våld. Hennes namn stod fortfarande kvar i sin lilla stålram. De vita plastbokstäverna avtecknade sig mot den mörkblå plyschen. Lundström. Så löjligt. Som om hon var en vanlig Lundström. En vanlig liten kvinna på väg hem efter en dags hårt jobb.

Li satte nycklarna i dörren. Vred om.

Ingenting rörde sig 2. dä_____ .

Hon öppnade dörren 3. fö_____ .

Bakom henne stod Nilla med dragen pistol.

I 4. ha_____ var det mörkt. Dagstidningen låg underst. Ovanpå den låg ett par räkningar. Ett kuvert från Försäkringskassan som avtecknade sig med sitt blåprickiga papper mot golvet. Ett vitt kuvert. Reklam från Libresse.

Li vände sig om och Nilla 5. ni_____ .

De 6. be_____ sig in i mörkret.

Det var så 7. t_____ att Katty hörde hur någon satte på kranen 8. i_____ hos grannen. En dörr smälde igen längre ner.

FACIT.

1. oskyldigt
2. därinne
3. försiktigt
4. hallen
5. nickade
6. begav
7. tyst
8. inne

Difference between produced tests in examples 1 and 2 is that the first one contains words of the same wordclass; whereas in the second case words of different wordclasses have been automatically selected from the same frequency band.

4.3 Computer-Assisted Generation of Multiple-Choice Items

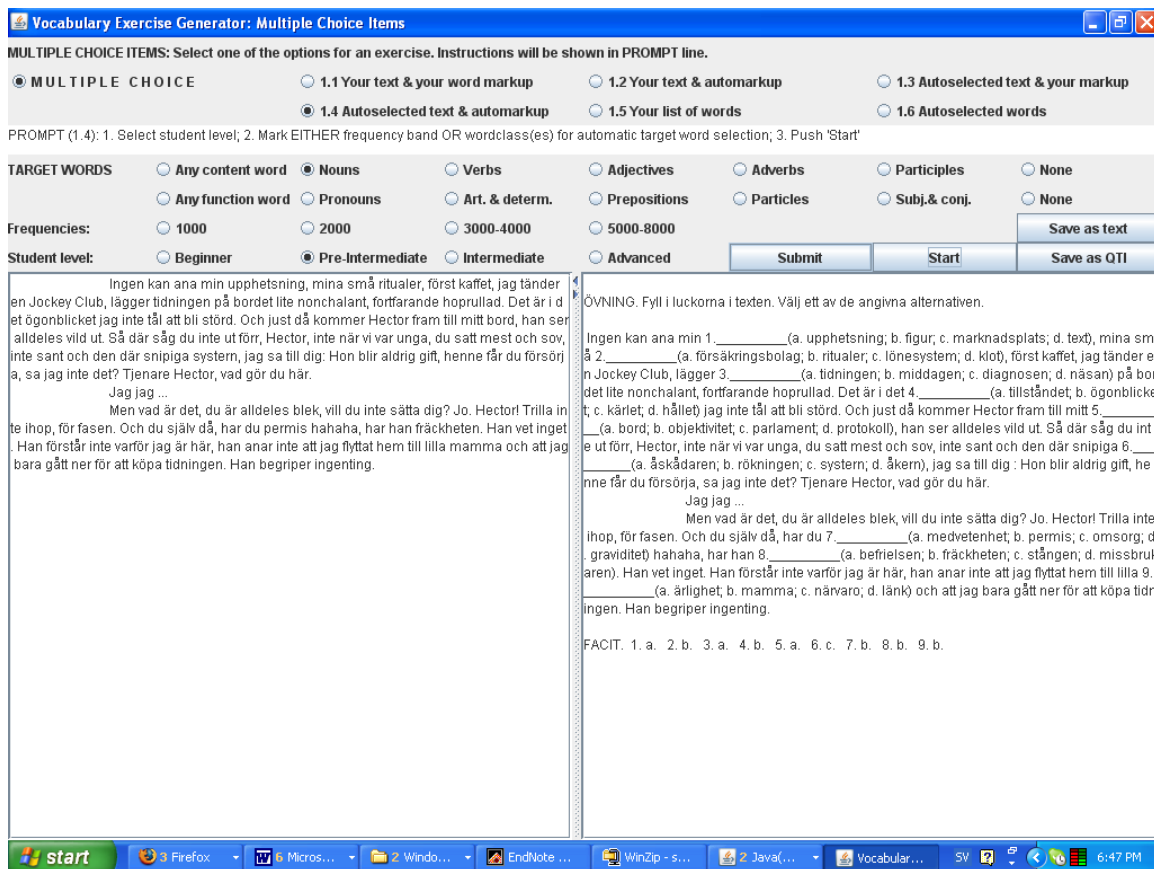
Multiple-choice items within vocabulary assessment have a long history and are commonly used even today in spite of the trend to use more embedded and contextualized ways of assessing vocabulary. The advantage of using multiple-choice items lies in their reliability, well-established and verified procedures as well as their consistent interpretability in terms of learners' vocabulary proficiency. At the same time researchers underline that item difficulty and its ability to indicate learners' level depend to a great degree on professionalism with which distractors are selected (Coniam 1997; Read 2000).

The issue of context also plays a significant role. It is normally recommended that at least a context of one sentence should be used to provide contextual clues for multiple-choice items. Passage-long and text-long contexts are also used in different tests. The use of decontextualized words, on the other hand, has been criticized.

Read (2000) names two main disadvantages of multiple-choice items: their construction is very time-consuming and their quality is too dependent on the choice of distractors. In this software I have tried to find ways to select distractors automatically, thus saving construction time. The quality of automatically generated items has to be evaluated under the real-life conditions later on. It is, however, obvious, that computer-assisted production of multiple-choice items, even accompanied by human filtering and proof-reading can save time for test-producers.

The options that this software offers as far as multiple-choice gapped items are concerned are as follows: the test producer can paste his/her text or let the program select text automatically from SUC. As far as selection of items for gaps is concerned it can be done either manually or automatically. For the latter the user needs to mark which frequency band should be tested. One more option allows the user to type/paste a list of words plus their word classes into the window letting the program select sentences from SUC automatically. The last option is to let the program select random target words for assessing from a specified frequency band, select sentences from SUC and construct an exercise automatically. Figure 12 shows how the authoring tool looks like and what options are offered:

Figure 12. Multiple Choice Module, user interface of the authoring tool



4.3.1 Selection of Distractors

There exist several approaches to selection of distractors for multiple-choice items, depending upon the purpose of testing. To take one example, in a knowledge test semantically related distractors are selected using WordNet facilities (Mitkov & Ha 2003). When applied to vocabulary training and testing, distractors may be selected according to:

- semantical closeness
- shared frequency band
- shared wordclass
- shared frequency band and wordclass
- closeness in orthography / phonetics (even homophones)
- definitions
- without any specified principle

(Aist 2001) quotes a different approach to distractor selection proposed by Nagy, Herman and Andersson (see in Aist, 2000, p.221):

- Level 1. Distractors are a different part of speech from the correct answer. For example, if the target word is *astronaut* and the correct answer is *traveler*, Level 1 distractors might be *eating*, *ancient*, and *happily*.
- Level 2. Distractors are the same part of speech but semantically quite different. For example, if the target word is *astronaut* and the correct answer is *traveler*, Level 2 distractors might be *antelope*, *mansion*, and *certainty*.
- Level 3. Distractors are semantically similar to the correct answer. For example, if the target word is *astronaut* and the correct answer is *traveler*, Level 3 distractors might be *doctor*, *lawyer*, and *president*.

With the NLP resources at hand, the option of semantically close distractors could not be applied since we have no access to Swedish WordNet yet. Orthographical and phonetic similarity has been also excluded from this program. The reason for the latter, however is the lack of time rather than lack of resources (e.g. Svenska Ord that contains phonetic information would allow us to select words on that basis). For gapped items shared frequency band and shared wordclass have therefore come as a natural choice. This approach to distractor selection is supported by a number of researchers (Coniam 1997; Read 2000; Brown et al. 2005).

It is, however, clear that adding semantic information to frequency information could have made the choice of distractors more sophisticated. Coniam (1997) has pointed out that though there is a clear relationship between word frequency and proficiency, it is still desirable to differentiate which sense of the word is used. (Graesser & Wisher 2001) have suggested guidelines for distractor selection which presuppose that distractors should have different degree of distractability. One of the distractors should be a “near miss” – very closely related to the correct answer and look seductive for the test-taker; the second distractor should be thematically connected to the topic, yet not correct; the third one, called by Graesser & Wisher “unrelated distracter”, can be unrelated to the text content.

The guidelines set up by Graesser & Wisher (2001) are difficult to follow when generating multiple-tests automatically. It could have been possible with more sophisticated NLP technologies available: to name one, NLP technology that would allow topical analysis of the text; another desirable resource would be semantic network for Swedish or Swedish thesaurus.

In this software the approach based on similar word frequency and grammatical form is assumed. Base vocabulary pool (Forsbom 2006) that is used for frequency information in this program contains information about lemma, its wordclass, and all morphological forms of the lemma with specified tags. Once having extracted the specified tag, it is a matter of simple lookup in the list of the same frequency to find words having the same tag (wordclass and grammatical form) and select them at random, see examples in Table 14:

Table 14. Examples of automatically selected distractors.

Target word	Frequency band	Specified tag	Selected distractors
systemet	1	.NCNSN@DS	ljuset, kravet, svaret
förklaringar	1	.NCUPN@IS	bakgrunder, anledningar, kontroller
pålitlig	5	.AQ	oundviklig, oförmögen, sur
på	1	.SPS	hos, med, genom

innehåller	1	.V@IPAS	skapar, tittar, faller
vidare	1	.RG	främst, naturligtvis, hemma
som	1	.CC	både, och, men

In rare cases when there is no enough distractors of the same FB, the closest FB is checked.

What is critical, though, is that when the test-producer marks words for gaps manually, the correct wordclasses are set. The system extracts a specified tag including morphosyntactic information from the frequency list. This is done automatically following the regular syntax used in the frequency lists.

It is of vital importance that the correct option should be genuinely correct (Alderson, Clapman & Wall 1995). This issue whether distractors selected on the above-described principles are appropriate or not needs testing in real life environment and is left for future work.

4.3.2 Selection of Sentences/Texts

According to Bormuth quoted in (Cedergren 1992) the same procedures can be applied to identify sentence difficulty level as to the texts. In this project we assume it as a hypothesis.

Text selection procedure is described in chapter 3.5 "SUC as a source of authentic examples". Sentences are selected either on the basis of provided words and wordclasses (user input) or on the basis of random word selection from a specified frequency band. The wordclasses that are allowed as user input in the program are quoted in chapter 3.5.3 "FL in selection of distractors for multiple-choice items and synonym items".

4.3.3 Scoring Procedures

The scoring assumed for multiple-choice items is based on one full point for every correct answer and zero points for incorrect ones.

4.3.4 Examples of automatically generated multiple-choice items

Multiple-choice items can be used for placement or diagnostic purposes as well as for lesson material training and for final tests. They can also be constructed based on frequency range, or on wordclasses; in text- or sentence-based formats. A list of words can, if desired, be fed as input to the program. Here are some examples:

Example 3. Multiple-choice items: automatic search for adverbs in a text of pre-intermediate level.

ÖVNING. Fyll i luckorna i texten. Välj ett av de angivna alternativen.

På det andra trädet föreföll äpplena mer rödglänsande och lockande. Arys kröp sakta 1. _____ (a. minst; b. utåt; c. härigenom; d. därmed) en gren, tills hon kunde nå dem med kniven och slå av dem med dess egg. De föll och häxan dansade av glädje nedanför.

" Det räcker ", sa hon. " Kom ner nu. Vi måste komma iväg. "

Hennes iver att komma bort från platsen skrämde Arys, som kommit att tänka på sin dröm. Hon släppte ner kroppen, hängde i händerna och släppte. Snön tog emot hennes böjda ben men hon föll ändå 2. _____ (a. därvid; b. därinne; c. omkull; d. någorlunda). Så högt hade hon aldrig hoppat förr. Hon blev

yr av fallet, av att snurra runt i snön, av stöten. 3. _____ (a. olika; b. synd; c. fortfarande; d. dessutom) bländade plötsligt snön och solen henne. Hon satte sig upp, kisade - och stelnade.

Arys ville ropa, men hennes hals slog knut på sig. Hon fick inte fram ett ljud.

Det var en mörk skugga därborta - men häxan såg den 4. _____ (a. fortfarande; b. alltmer; c. inte; d. likaväl), för hon hade fallit på knä och skrapade med händerna i den hårdpackade snön. Skuggan såg först ut som en stor man, men Arys förstod snabbt vem den var. Den var 5. _____ (a. plötsligt; b. drygt; c. korrekt; d. internationellt) på väg mot häxan, mycket snabbt. 6. _____ (a. därav; b. gratis; c. då; d. precis) lossnade ljuden i flickans hals. Hon skrek ut en varning och trevade förtvivlat runt sig i snön efter kniven.

FACIT. 1. b. 2. c. 3. d. 4. c. 5. a. 6. c.

Example 4. Multiple-choice items: automatically selected nouns for training in sentences of intermediate level.

ÖVNING. Fyll i luckorna i meningar. Välj ett av de angivna alternativen.

1. Vi kör med _____ (a. solidariteten; b. satsen; c. vakten; d. målsättningen) att inte öka lagret och just nu har vi balans i produktionen .

2. Arne Lanning hade levt bortglömd av _____ (a. lukten; b. pressen; c. fienden; d. skivan) alltför länge .

3. Kollegan som fortfarande stod bakom Robert grymtade fram någonting som med god vilja kunde tas för ett _____ (a. samspel; b. flöde; c. trä; d. skratt) .

4. Vår identitet kan ses som dels _____ (a. riksdagen; b. summan; c. ersättningen; d. båten) av alla dessa roller - vår totala rollrepertoar - dels just den förmåga (eller oförmåga ibland) vi har att hantera de här olika rollerna i olika situationer .

5. Då släppte Robert _____ (a. beslutet; b. materialet; c. taget; d. stället) och den andre stod flämtande kvar på knä och gned kvidande sin onda handled .

6. * Påverka _____ (a. u-ländernas; b. hjulens; c. sågverkens; d. myntens) ekonomiska politik i riktning mot ökad privatisering och marknadsprissättning , mot konvertibla och rimligt värderade valutor och mot tillskapande av lagar och andra affärsjuridiska regelsystem för industri och handel .

FACIT 1. 1. d. 2. b. 3. d. 4. b. 5. c. 6. a.

FACIT 2. 1. målsättningen 2. pressen 3. skratt 4. summan 5. taget 6. u-ländernas

It would be both interesting and useful to test different automatically created items in the real-life conditions or at least to ask experienced test item constructors to evaluate the quality of automatically created items. It might happen that certain types of exercises are more useful than others. Coniam (Coniam 1997) describes an experiment with a system that could create multiple-choice items of different types. Test-items with every *nth* deleted word were found less acceptable than the two language oriented modes of selecting words from a specified word frequency band and particular wordclasses.

4.4 Computer-Assisted Generation of Word Bank Items

4.4.1 General Information on Word Bank Items

The last format of gap items implemented in SCORVEX is based on the principle of collecting all extracted words in a list offering to choose the most appropriate alternative for each gap. This technique is widely used in L2 assessing. Several variations can be observed:

- selecting target words without any system;
- to complicate the task for the test-taker all extracted words can be of the same part of speech. This way the student will not use any other clues than lexical for choosing an appropriate word;
- yet another way to complicate the task for the testee is to offer more choices than there are gaps. This way the student cannot merely guess which word goes into which gap, inserting the ones that he/she knows and leaving more difficult till the end. He or she has to be more critical in choosing the correct alternative;
- finally, the learner might get a task of inserting uninflected words into appropriate gaps, putting them into grammatically correct form.

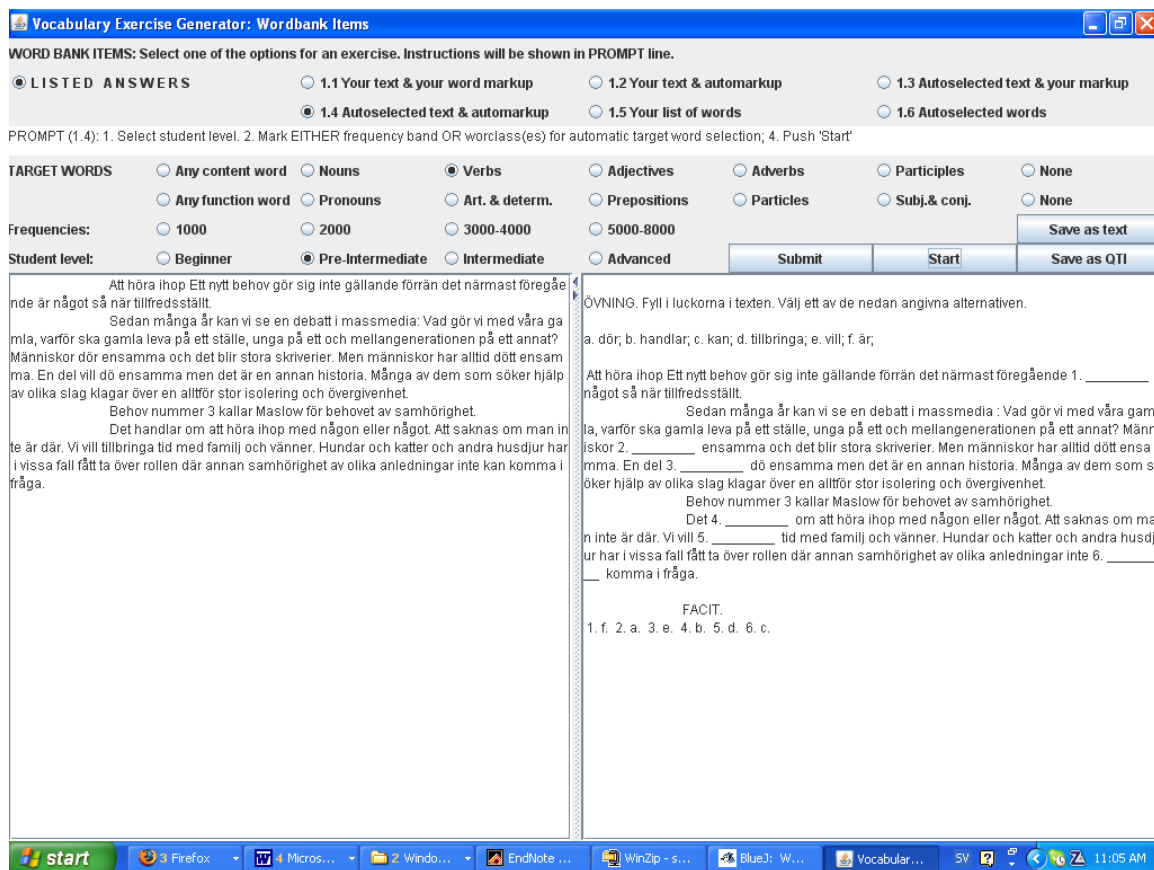
In this software we are following the first and the second principle, i.e. selecting any words, and leaving equally many words in a wordbank as there gaps. With annotations available in SUC it will not be very difficult to alter the program so that any other alternatives are used instead.

As with multiple-choice format, wordbank vocabulary items can be generated from texts and wordlists, and the markup of target words can be either done manually or automatically with reference to the frequency band or specified wordclass. To avoid having gaps following close one after another a constraint is used assuring that gaps can be at a distance of minimum 5 words.

Calculation of scores is based on one full point for each correct answer and zero for each incorrect one.

The created exercise looks as follows in the user interface window (Figure 13):

Figure 13. User Interface of the Word Bank Items Module



4.4.2 Examples of automatically generated word bank items

Wordbank items are the easiest exercise type for construction. Yet, they allow varied training of vocabulary. Examples 5 and 6 show exercises for training different forms of the same lemma:

Example 5. Word bank items: exercise created on the base of a list of manually typed words (5 times the same word). Variant 1

ÖVNING. Fyll i luckorna i texten. Välj ett av de nedan angivna alternativen.

a. vackraste b. vackert c. vackra d. vacker e. vacker

1. Santa Marias tårar , tänkte Katty och slängde sig efter dem , hon jagade dem bland dammråttorna under sängen , kröp efter dem in under garderoben och förde dem tillsammans igen , försiktigt , med handflatan , som en grupp busiga barn som inte ville stå still och sprang åt alla håll , tills de blev orörliga och utgjorde en _____ liten pyramid vid hennes fötter .
2. Abdel Gamal var oerhört stolt över sina vita dromedarer , de ädlaste , snabbaste och _____ av öknens alla djur , han hade förtjust lyssnat till lovsångerna .
3. Jag hade nog vågat hoppas att hon - som jag tidigare sett på bio - skulle ge sig ut på en kort plingande slädfärd i det _____ vintervädret .
4. Efteråt kommer min dotter upp till mig på åskådarläktaren , en inte alltför _____ syn : hon haltar efter en smäll på knät i matchen mot spanjorskan och har en rejäl fläskläpp , förutom en del mindre skrämor i ansiktet .
5. - Det är så _____ här .

FACIT. 1. e. 2. a. 3. c. 4. d. 5. b.

Disadvantage of this sort of item is that the same wordform appears more than once in the keys. So far this has not been taken into consideration for calculation of the score. If the student chooses alternative "d" (vacker) where it has been saved by the program "e" (vacker) as the right answer, this will be calculated as an error. The problem is, though, a minor one and can be solved easily in the future.

To avoid the above-mentioned problem and simultaneously make this type of items more difficult it is possible to ask the student to put the uninflected form of the word into the necessary form, as shown in example 6:

Example 6. Word bank items: exercise created on the base of a list of manually typed words (5 times the same word). Variant 2.

ÖVNING. Fyll i luckorna i texten. Använd rätt form av ordet BORD

1. Hon stod vid huvudändan av det långa _____ och väntade på att de skulle ta plats kring det , men ingen ville tydligen sitta där , istället hade man samlats i grupper kring väggarna där de äldsta och kvinnorna med småbarn slagit sig ner på de läderklädda bänkarna .
2. Däremot lät de blickarna svepa över allt annat : _____ och stolar , flugfångaren över spisen , väggfotogenlampan med sin mässingsreflektor , vägghyllan med vågen , den sällan använda kaffekvarnen och burkarna med mjöl , socker och salt .
3. Vi har nu ett bildschema för vart och ett av orden i satsen " Lampan är över _____ . "
4. Det var då allt det här kom till : rosenplanteringarna , de nya uthusen med rum för gästande sökare , bersåerna , de små runda _____ där man kan sitta och dricka likör .
5. Man erbjuds att ta plats vid ett av _____ och sedan lämnas man ifred , om man inte själv tar kontakt .

FACIT. 1.bordet; 2.bord; 3.bordet; 4.borden; 5.borden;

Such items can even be used to introduce new vocabulary and demonstrate contexts where the new word can be used.

Another possible area of application of vocabulary items of the wordbank type is differentiating between frequently confused words, as shown in examples 7 and 8:

Example 7. Word bank items: differentiating between different forms of pronouns. Target vocabulary has been typed in by the user (not automatically generated!)

ÖVNING. Fyll i luckorna i texten. Välj ett av de nedan angivna alternativen.

a. som; b. vilket; c. vilken; d. vilka;

1. Det finns också förslag till ett EG-direktiv om skydd för gravida eller ammande kvinnor , _____ lägger starka restriktioner för arbete med misstänkt fosterskadande ämnen , bland dem bly .
2. Kommunikationsmedlen styrde reseströmmarna till bestämda stråk , från _____ man sällan avvek .
3. Professor Daniel Callahan , rektor för The Hastings Center , ett ledande bioetiskt centrum i USA , gav i slutet av åttiotalet ut en bok , Setting Limits , _____ väckte en häftig debatt .
4. Kostbara livsmedel vägdes på skålvågar _____ gav namn åt viktenheten skålpund .

FACIT. 1. a. 2. d. 3. c. 4. b.

Example 8. Word bank items: differentiating between different forms of participles. Target vocabulary has been typed in by the user (not automatically generated!)

ÖVNING. Fyll i luckorna i texten. Välj ett av de nedan angivna alternativen.

a. störande; b. störd;

1. Men förutom det glädjande meddelandet innehöll brevet också flera _____ och sårande inslag , som fick henne att börja minnas sådant som hon hoppats slippa tänka på igen .
2. Men sant är också att man blir _____ !

FACIT. 1. a. 2. b.

Items of the above-mentioned types (examples 5 – 8) are suitable as progress tests or for training of lesson materials. Vocabulary items based on automatic selection of target vocabulary from a particular frequency band, on the other hand, are more suitable for diagnostic or placement tests or as final tests. The items of the latter type can be generated either in the form of sentences or as a text, as shown in examples 9 and 10:

Example 9. Word bank items: automatically selected words from FB2

ÖVNING. Fyll i luckorna i texten. Välj ett av de nedan angivna alternativen.

a. utrymme; b. bolaget; c. knä; d. ifrån; e. förmår; f. föreskrifter;

1. Deras stuga låg för nära gruvan , deras tillstånd och tillmötesgående krävdes innan _____ kunde börja sin brytning .
2. Hebréerbrevet förklarar visserligen att den judiska lagens _____ om ren och oren mat inte bör tolkas bokstavligt .
3. Du får nöja dig med vad huset _____ .
4. Bengt Nilsson gav _____ sig ett frustande läte .
5. Han satte sig på _____ igen , böjde sig ner för att lossa gallret och lyfta upp Samantha .
6. Ett förtjusande _____ , kallat " the study " , är min privata toalett .

FACIT. 1. b. 2. f. 3. e. 4. d. 5. c. 6. a.

Example 10. Word bank items: automatically selected text for level 3 with automatically marked words from FB3

ÖVNING. Fyll i luckorna i texten. Välj ett av de nedan angivna alternativen.

a. bara; b. begriper; c. blek; d. fasen; e. försörja; f. gift; g. störd; h. tänder; i. vild;

Ingen kan ana min upphetsning, mina små ritualer, först kaffet, jag 1. _____ en Jockey Club, lägger tidningen på bordet lite nonchalant, fortfarande hoprullad. Det är i det ögonblicket jag inte tål att bli 2. _____. Och just då kommer Hector fram till mitt bord, han ser alldeles 3. _____ ut. Så där såg du inte ut förr, Hector, inte när vi var unga, du satt mest och sov, inte sant och den där snipiga systemen, jag sa till dig : Hon blir aldrig 4. _____ , henne får du 5. _____ , sa jag inte det? Tjenare Hector, vad gör du här.

Jag jag ...

Men vad är det, du är alldeles 6. _____ , vill du inte sätta dig? Jo. Hector! Trilla inte ihop, för 7. _____. Och du själv då, har du permis hahaha, har han fräckheten. Han vet inget. Han förstår inte varför jag är här, han anar inte att jag flyttat hem till lilla mamma och att jag 8. _____ gått ner för att köpa tidningen. Han 9. _____ ingenting.

FACIT. 1. h. 2. g. 3. i. 4. f. 5. e. 6. c. 7. d. 8. a. 9. b.

Choosing words of the same wordclass for training excludes “guessing” strategies based on syntactic context. Instead, a student applies his or her knowledge of the word through analysis of the context where the item can fit semantically. A further refinement can be made by adding one more alternative than there are gaps; or by providing an alternative that says “wrong word”.

Training words of a particular part of speech (within a particular frequency band/or in any band), can also be made text- or sentence-based, as demonstrated in examples 11 and 12. These items, too, are most suitable for final tests or as diagnostic/placements tests.

Example 11. Word bank items: automatically selected prepositions for training in sentences

ÖVNING. Fyll i luckorna i texten. Välj ett av de nedan angivna alternativen.

a. på; b. mellan; c. per; d. för; e. innanför; f. rörande;

1. - De yttersta tålederna reflekterar hjärnan och på höger fot behandlas vänster hjärnhalva , som står _____ det rationella , rätlinjiga tänkandet men även för katastrofkänslor .
 2. Hon står fortfarande precis _____ dörren och de kan inte se varandra när de pratar .
 3. Det finns en traditionell semantisk teori som säger att ett ords betydelse bestäms av en koppling _____ språket och tingen i världen .
 4. Nästa år får man spara ända upp till 1200 kr _____ månad sammanlagt på en av de två allemanssparformerna .
 5. Även halshuggning var en offentlig förrättning _____ torget .
 6. Carl var på nytt den uppenbart tillfrågade och han gav sig in på en försiktig utläggning _____ kända eller förväntade motsättningar i den ryska statsledningen och vilken framtida betydelse detta kunde få .
- FACIT. 1. d. 2. e. 3. b. 4. c. 5. a. 6. f.

Example 12. Word bank items: automatically selected prepositions in an automatically selected text

ÖVNING. Fyll i luckorna i texten. Välj ett av de nedan angivna alternativen.

a. av; b. hos; c. i; d. med; e. mot; f. på; g. till;

- Britt och jag såg vad som gömdes 1. _____ ritningarna, fortsatte han och rösten blev allt svagare. Tala 2. _____ din pappa. Säg " sub terra ". Han kommer att förstå.

- Vad betyder det? undrade hon.

- Sub terra ... Rummen. Sub terra. Professorn kommer att förstå, viskade han utmattad. ST.

Förstora ritningen!

Hon lade handen 3. _____ hans kind och sade hans namn. Han svarade inte. Doktor Dotvic kom när hon ropade, och han kunde bara konstatera att Silver gått in i koma igen.

- Vi vet inte när han vaknar igen. Det dröjer säkert några timmar.

- Jag stannar 4. _____ honom, sade hon och läkaren märkte hur hennes ögon tårades.

- Du ska inte stanna. En mördare finns där ute. Se till att ta fast honom i stället. Staffan har ingen glädje 5. _____ att du stannar. Det är bättre att du får tag 6. _____ den som såg till att han hamnade här.

- Han kommer att klara sig, sa hon, mest 7. _____ sig själv.

FACIT. 1. c. 2. d. 3. e. 4. b. 5. a. 6. f. 7. g.

Clearly, certain human proofreading and testing is needed before estimating the degree to which the generator of c-tests, multiple-choice items and word bank items can be of use. Certain generated items might need to be corrected or even discarded; the amendment facilities that are lacking in the present version of SCORVEX are planned to be implemented into the software in the future.

However, demonstrated examples of automatically generated exercises throw light on the advantages of using corpora as a source of teaching material. Even though the program cannot identify the learner competence in language skills automatically, it can select texts of appropriate difficulty if the test creator marks the right student level. The rest can be handled totally automatically without human intervention. It is thus possible to create teaching material in a matter of seconds and cover the needs of a homogeneous student group with different proficiency levels. This vocabulary exercise generator can therefore become an effective tool for generating an infinite number of items for variable proficiency levels.

4.5 Swedish Vocabulary Size Test

4.5.1 General Information on the Test Design

One of the crucial issues for vocabulary knowledge testing is the evaluation of a learner's vocabulary size. It can be done in several ways. One of them is by means of yes-no test, a test that measures learners' passive vocabulary knowledge. Swedish Vocabulary Size Test is an example of a "yes-no" test.

Eurocentres Vocabulary Size Test, the prototype for the Swedish Vocabulary Size Test described in this section, is a test designed both as a placement instrument demanding minimum efforts from administration and as an instrument for measuring a learner's vocabulary size. The test is extremely time-saving: it can be done in less than 15 minutes by a learner and provide test organizers with reliable results on the spot.

The test was originally designed at Birkbeck College, University of London by Paul Meara and his colleagues (Read 2000). It was accepted by Eurocentres, an organization providing courses in English in many countries, as a placement instrument. Eurocentres needed a quick and efficient placement procedure for assigning students to different level groups with minimum administrative efforts. This test met both requirements: it could be taken in about 15 minutes, was administered by a computer, thus saving time on correction work, and provided immediate results giving good ground for dividing students into level groups. The validity of this kind of placement is arguable; in case it is used it is based on the assumption that vocabulary knowledge is central for language proficiency in general, and in particular that the number of words a learner knows can indicate his/her language proficiency.

The test is formed as a checklist consisting of words from numerous frequency bands and a proportionally large amount of nonsense words. For each frequency band there are 60 test items: 40 existent and 20 non-existent words (Huitbregtse, Admiraal & Meara 2002). The non-existent items are used to adjust students' scores in case they tend to overestimate their knowledge. Learners are warned that a certain amount of test items are non-existent words and are afterwards presented with a question: "Do you know this word?". Learners have to answer the question by clicking "yes" or "no" button accompanying each test item. The score is reduced if a learner claims that he/she knows some pseudowords.

There has been a lot of discussion about the validity of the test. Some experiments have shown that the test lives up to the purposes when compared to other placement instruments (see references to Meara & Jones, 1988, in Read, 2000, pp.127-128). Vocabulary Size Test gives some percentage of misplacement, but so do a lot of other placement instruments. It is, however, underlined that a pure vocabulary size test should be complemented by other placement procedures, e.g. interviews, grammar tests, etc. to give more objective placement information.

Pedagogically viewed it is a yes-no test that measures learners' receptive knowledge of vocabulary, i.e. words outside of language context. It has its advantages and disadvantages but we leave this discussion outside this work. More information about the validity of the test can be obtained from Read (2000). The Eurocentres Vocabulary Size

Test has been taken as an inspiring example for this module of the exercise generator and has been adjusted to the Swedish language. In its present form it can be used as a placement/diagnostic test or vocabulary size test.

There are, however, some differences in the way the original test for English and the test for Swedish are designed. First, the vocabulary size test for Swedish can be automatically generated for eight frequency bands and demands no manual test construction. It can be saved in a paper variant or in QTI format for later use, taking less than a couple of seconds to generate a new test for each frequency band. Second, SCORVEX generates tests for 8 frequency bands instead of 10, using an adjusted frequency list derived from Stockholm Umeå Corpus (SUC) (Forsbom 2006). Third, the potential words are automatically generated on a different principle (see sub-section “Generation of Potential Swedish Words”).

4.5.2 Generation of Potential Swedish Words

Nonsense words or pseudowords, as they are called by Meara and his colleagues (Huitbregtse et al. 2002), are words that fulfill phonotactic constraints of a target language, but which are not present in the language system. In Eurocentres Vocabulary Size Test the principle mechanism for coining pseudowords is through combining existent syllables specific for the tested frequency bands into new words. This is done manually.

In the Swedish vocabulary size test another approach has been used. The starting point has become the syllable structure of Swedish words generally presented as

$$(C(C(C))) V (C(C(C)))$$

where C is a consonant and V is a vowel. There can be zero, one, two or three consonants at the beginning of the word, a vowel that is the only obligatory part of a syllable and zero, one, two or three consonants in the end. The final consonant clusters can in fact be longer than three consonants, for example “västkustskt”. Clusters longer than three consonants account for the marginal cases.

The phonotactic structure of Swedish syllables including initial, medial and final consonant clusters as well as possible combinations between consonants and vowels are described in (Sigurd 1965) and (Elert 1970). Following their descriptions a number of central one-, two- and three-consonant clusters have been selected, constraints between consonant clusters and vowels described, and a number of central final consonant clusters added to the program. Marginal cases have been dismissed as well as a possibility of coining longer root morphemes. The latter is due to the fact that average syllable length of different frequency bands lies within three-syllable words. Pseudowords that are generated as test items have to be of the average size, plus/minus one syllable. Appendix 5 contains all initial and final clusters used in the program as well as suffixes and prefixes used for generation of potential words.

A program can, according to the rules and constraints described in the program, combine initial consonant clusters with vowels and final clusters, thus coining a one-syllable root morpheme. The resulting pseudoword is checked against a lexicon database and if there

is no such entry in the database, the assumption is made that the word does not exist in Swedish and therefore it is added to the list of test items.

In case the program asks for a longer word, which is the case with frequency bands over the threshold of 4000 where first three-morpheme words are coined, a number of existing suffixes or prefixes can be added to the coined root morphemes. This is a reasonable simplification of coinage and phonotactics for the purposes of this test. It has proved to create words that sound and look more like Swedish words with derivational affixes than coined two or three syllable words without affixes. Some examples of pseudowords:

dri, rylt, krämb, vräpt, jench, fov, spjägande, ingtlig, utman, läism, kvingdisk, träldant, splaving, späare, uhet, hyrthet, eant, inelse, bäbar, gråkhet.

Initial consonants, vowels and final consonants are selected and combined at random following certain phonotactic constraints described in the program. Suffixes and prefixes to be added follow the same principle of random selection, filtering certain three-consonant clusters or unacceptable consonant combinations, e.g. root morpheme “sass” + suffix “tion”. The validity of resulting pseudowords has to be tested by some learners of Swedish.

It can unfortunately happen that words that are coined have the form of an existing inflected lemma, e.g. “täckt”, but have not been sorted away by the program due to the fact that the lexicon database against which pseudowords are tested does not contain inflected forms. This problem has to be addressed in the future.

Discussion about the usefulness of pseudowords in “yes-no” test can be found in Read (2000).

Had there been a publicly available and reliable syllable parser for Swedish, another mechanism for coining pseudowords could have been used. It would be possible to parse words in each frequency band for syllables and join initial, medial and final syllables together to get pseudowords for the particular frequency band test.

4.5.3 Calculation of the score

Looking simple and easy to handle on the surface, this test presents in fact a number of interesting questions when it comes to scoring procedures.

If a test-taker marks everything with “yes”, he or she can score 40 points out of 60. It will say little about the vocabulary knowledge of the learner. It will, on the other hand, reveal a lot about the test-taker’s guessing-strategy and his/her response style.

Huibregtse et al. (2002) discuss four different scoring algorithms. The starting point for their discussion is the response alternatives which are presented in a table borrowed from their article (Figure 14):

Figure 14. Stimulus-response matrix taken from (Huitbregtse et al. 2002)

		Response alternative	
		Y	N
Stimulus alternative	w	$P(Y w)$ hit	$P(N w)$ miss
	p	$P(Y p)$ false alarm	$P(N p)$ correct rejection

Figure 1 Stimulus–response matrix

Notes: The stimulus alternatives are w for word and p for pseudoword. The response alternatives are yes (Y) and no (N).

They argue that calculating all correct answers (“yes” for real words and “no” for pseudowords) seems to be too simple. The scoring procedure should take into account a number of variables: beside vocabulary knowledge itself, guessing and response style influence the result. Response style is characterized by learners’ approach to a word that they are not sure of. Some learners tend to say “yes” if they are in doubt, others tend to say “no” in the same situation. Both tendencies should be compensated in the final score.

The four scoring procedures are discussed:

- the number of correct responses – the easiest and seemingly most obvious way of calculating the score which, however, does not take into account test-takers’ response styles;
- correction for guessing – the assumption here is that every false alarm (saying “yes” to pseudoword) is a result of blind guessing and the score should therefore be recalculated accordingly. Huitbregtse et al. (2002) argue that this calculation method does not take into account response style and views knowledge of a word as either perfect (100% knowledge) or absent (zero knowledge);
- signal detection theory: Meara’s Δm – the calculation of the score is based on statistics from signal detection theory, developed originally for military purposes. The theory estimates two aspects of human performance: the ability to discriminate and the response bias. The way Meara has used Δm , it is argued in the article, that the scoring does not correct for individual response style in an adequate way;

- signal detection theory: a new index. A corrected variant of Meara's Δm . The formula is based on statistic values for hits and false alarms and their interrelation. For more details see Huibregtse et al. (2002).

In this test the last scoring algorithm is used. It has its advantages and disadvantages. The advantage is that it takes into account the response style of each learner. If the learner is hesitant and tends to say "no" in all doubtful cases, his/her scores are not reduced in case there are no false alarms. If, on the other hand, a learner tends to say "yes" in most cases, the "guessing" is compensated if there are false alarms. The disadvantage is that in case the hits are proportional to false alarms, the final score might be "0". However, this does not mean the learner has no vocabulary knowledge at all but rather that the test responses do not provide enough ground for meaningful estimation of the learner's vocabulary knowledge.

The formula for this calculation is the following (comes from Huibregtse et al., 2002, p.238):

$$I_{SDT} = 1 - \frac{4h(1-f) - 2(h-f)(1+h-f)}{4h(1-f) - (h-f)(1+h-f)}$$

where I_{SDT} stands for Index SDT (an algorithm for calculating the score);

f = false alarms;

h = hits;

User interface of this module is shown in Figure 15.

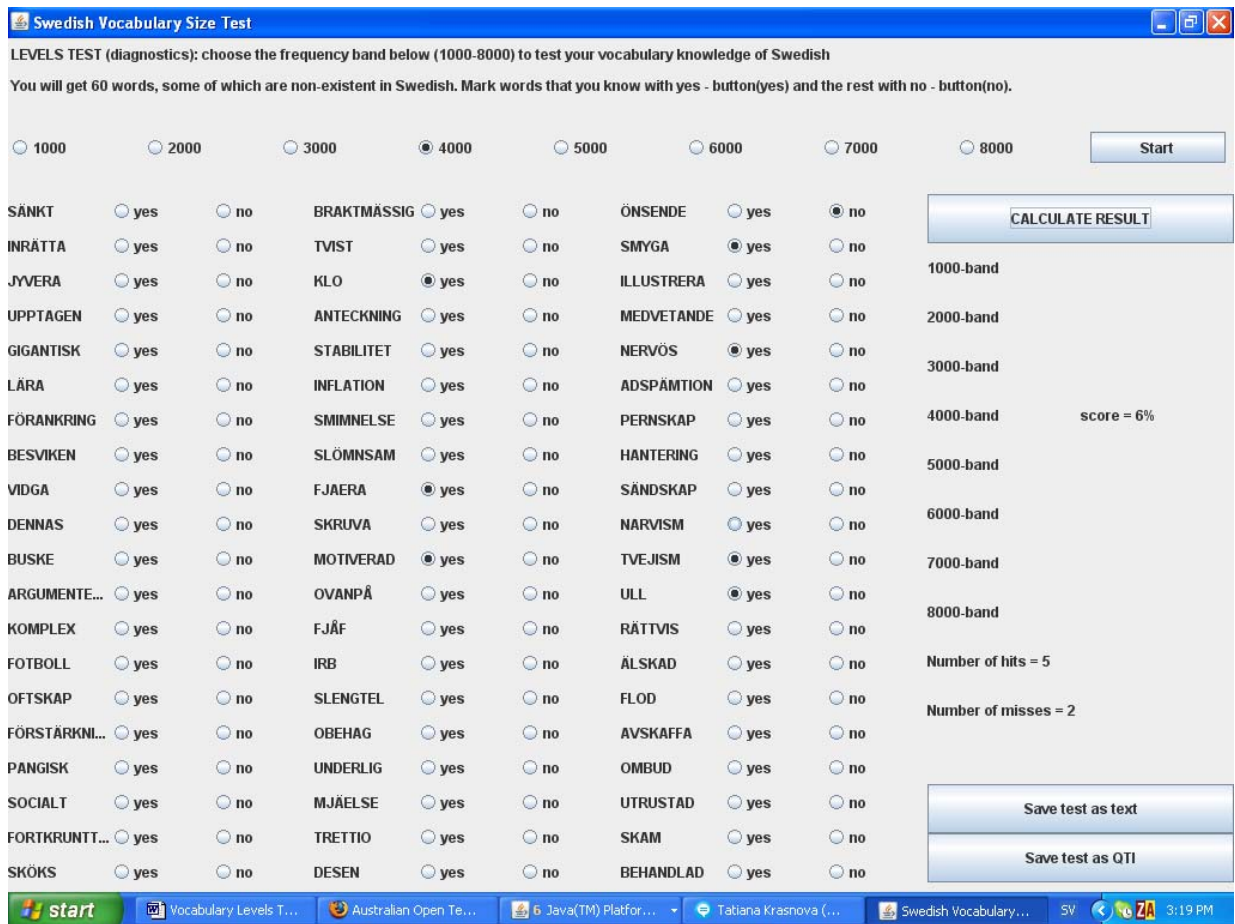


Figure 15. Swedish Total Vocabulary Test – User interface

5. Concluding remarks

In this concluding chapter I am naming the advantages and disadvantages of SUC for automatic generation of teaching materials, briefly outline how SCORVEX can be further developed and improved, and summarize the results of the research.

5.1 SUC – Advantages and Disadvantages

Before I start enumerating advantages and disadvantages of using SUC in automatic generation of exercises, I would like to say a few words about some issues that have either been solved in this generator or belong to discussable aspects.

Many corpus linguists claim that any corpora of written language below 5 million words are considered small corpora. SUC, that comprises 1,2 million running words, is therefore considered a small corpus according to some researchers (McEnery & Wilson 2001; Hunston 2002; O'Keeffe et al. 2007). It is, however, the design of a corpus that makes it appropriate or inappropriate for the planned activity. Type of texts, corpus structure, kind of annotation are therefore more relevant factors in terms of suitability as opposed to the size of a corpus. For language learning purposes, especially for automatic generation of exercises, there is no need in a huge corpus. A corpora of well-selected texts of about 1 million words is enough to provide varied texts and examples for pedagogical aims and needs and can thus be a source of valuable teaching aid (Dodd 1997; O'Keeffe & Farr 2003). In this respect SUC is a reliable, balanced corpus that lives up to the purposes of the generator and its size is clearly satisfactory for the needs.

Another aspect that is often taken up by linguists is that corpora used in teaching languages need to have specific design and texts appropriate for learners of different levels:

It seems that, ideally, texts for a CALL database ought to be pre-selected from a corpus with great discrimination. Teachers should be satisfied that all texts are models of good practice in word usage, syntactic constructions and cohesive discourse. Texts must be graded according to readability but must also be classified by distribution of linguistic features to ensure that the database coverage is adequate for the task. (Wilson, 1997, p.130)

An important issue is raised here – appropriateness of texts for learners and readability analysis of texts. In the course of work I have realized that the most important problem for the generated exercises have been automatically selected texts and sentences of inappropriate difficulty. Exercises that are generated on the basis of inappropriate texts are clearly unsuitable for pedagogical application. Luckily, this issue has been solved in this generator, as described in subchapter 3.5, but only with respect to lexical complexity and general readability measures. If the system is extended to cover even grammar exercises in the future it might become important to analyze texts as far as the distribution of grammatical structures is concerned, especially with respect to which grammar is taught at which proficiency level.

Accessing frequency data obtained from SUC is the pre-requisite for this generator, though there is one drawback connected with that. Disadvantage in using frequency lists based on SUC lies in the fact that they are based on written language in general whereas

the aim of language teaching tends to concentrate on conversational fluency, which in its turn demands some other vocabulary and grammar constructions as its core. The aim of language teaching is to combine conversation skills (speaking and listening) with writing and reading skills. Frequencies based on SUC are optimal for the written mode of language, conversational skills being left to teachers' intuition rather than statistical approach to vocabulary selection. Originally the idea has been to complement written language frequencies with spoken language frequencies derived from GSLC (Gothenburg Spoken Language Corpus), but has been abandoned due to the lack of time.

Advantages of SUC for automatic production of vocabulary exercises:

- availability for academic purposes with individual license;
- representativity: texts come from different genres, areas and topics and represent different proficiency levels;
- size: randomly chosen examples, sentences and texts do not have a risk of being repeated;
- possibilities it provides – frequency information, statistics of different kinds;
- annotation: there are a number of features that are annotated and make it possible to define search parameters acc. to the needs and thus to select examples (texts and sentences) according to the desirable inquiry automatically.
- variability within the lexeme (lemma) is caught through the annotation of lemma (base form with its word class) for each word in the corpus. This makes selection of desirable target items much easier; in addition to that it is possible to provide a lemma as a search criteria and receive examples with the word in different forms;

Disadvantages:

- absence of semantic information and absence of tools that can classify selected examples according to different meanings. Searching for example sentences with the word '*drottning*' (Eng. 'queen') can result in a list of sentences where '*drottning*' is used as a political figure as well as a kind of a bee or a chess piece. To select examples with the necessary meaning of the word will require manual disambiguation by the teacher or a special NLP tool that will be able to group examples according to their meanings. It is also desirable to have an NLP tool that can immediately, on a search query, inform the user which sense of the marked word is most frequent in the corpora, and which context is the most representative so that the examples used by the teacher are pedagogically appropriate.
- no disambiguation of homographs available. Homography between parts of speech has been dealt with through grammatical annotation; however, homography within the same part of speech is not disambiguated; for instance the entry *sticka* (verb) includes two different verbs - *sticka* (verb, *stickade*) och *sticka* (verb, *stick-stuckit*). Their frequencies are counted together as the same lemma/word. To distinguish between them and get reliable frequency information

each case needs to be analyzed individually. There are even several missing forms (in brackets I have given the number of occurrences of the word-form in SUC/Parole - “stickade”(0/39), “stickades”(0/0), “stickats”(0/0), “stuckits”(0/6), “stucken”(0/9), “stuckna”(0/3) - as the example below shows

```
1524  sticka.V      50.712288   6      stack.V@IIAS.41
      stacks.V@IISS.1  sticker.V@IPAS.25  stickar.V@IPAS.1
      sticks.V@IPSS.1  stuckit.V@IUAS.6  stickat.V@IUAS.2
      stick.V@M0AS.2   sticka.V@N0AS.9
```

Desirable features:

- annotation of syntactic functions (subject, object, etc) and phrase-structure (noun phrases, verb phrases, etc.). Having these features could facilitate generation of grammatical exercises for e.g. word order, which otherwise can be impossible to generate automatically. The syntactic annotation of SUC is being done at Uppsala University.
- annotation providing information on what syntactic structures words can enter, e.g. verbs followed by direct object or particles etc. Exercises could be made more varied.
- discoursal and text linguistic annotation;
- style tags (colloquial words, bookish, etc.);
- key words mark-up, e.g. words specific for certain topics, like architecture, politics, etc. This could facilitate selection of texts according to the student interests and to automatically identify topic of each text.
- annotation of idioms and other collocations;
- annotation of morphological constituents of every word, e.g. root morpheme, affixes, etc. This could facilitate grouping words into word families and generation of exercises on word-building.
- annotation of verb groups (1, 2, 3, 4th group) and noun groups for training any particular paradigm.

The wish-list can be extended, but the above-mentioned aspects could have definitely helped creating more “intelligent” and more varied types of exercises based on SUC.

5.2 Future of SCORVEX

SCORVEX is in no way a complete or perfect system. The existing modules can be improved in a lot of ways; there are still bugs and unsolved questions. The system can be expanded on the lexical level and other than lexical types of exercises can be built into it; user interface can be made user friendlier and techniques on how to present generated teaching materials need to be analyzed and implemented. Moreover, the generator needs to be tested in real-life environment to identify flaws and weaknesses. Thus, there are five main directions in which the system can be further developed, which are described below.

5.2.1 Towards the specificity of existing exercises

The way SCORVEX is built now, it is possible to make exercises more specific: for example, to introduce an option for generating exercises for training agreement between adjectives and nouns in noun phrases. This could be made possible via multi-tag search, i.e. looking for adjoining nouns and adjectives.

More sophisticated variants of existing exercises can be generated – for example with more answer choices than there are gaps. Another option is to provide uninflected base forms that should be inflected before they are used in sentences/texts (in case word bank items or multiple-choice items are used). This way a more advanced training is possible and grammatical clues are excluded.

Another way of enriching the system is to add search possibilities for creating exercises for distinguishing words, e.g. *vilken/vilket/vilka* versus *som*, or some others that cause learner confusion. Search could be then defined in terms of how many examples of one word versus another should be found. Even at present this is possible when providing user's own list of words in word bank items module with words repeated *n* times if more than one example with them is desired.

Access to some PoS-tagger and lemmatizer can facilitate part-of-speech analysis of user texts so that no part of speech confusion arises when e.g. searching for words of a certain frequency or for words of a certain wordclass. As it happens now, homonymy between wordclasses is neglected (when any user text is pasted into the interface) and words of one wordclass can be assumed to be a different lemma.

If the software is used for creating vocabulary items on other texts than the ones from SUC, than certain other NLP tools are necessary for more intelligent analysis of input texts, among other things PoS-tagger and lemmatizer. At the moment of implementation there was no available tagger or lemmatizer that could be imbedded into the software on a plug-and-play principle. The work with texts that are not extracted from SUC database is therefore based on naïve principles. When identifying items from a certain frequency band or of a particular wordclass in a non-SUC text, text words are matched against frequency lists. Due to their unique entries it is possible to find matches for both inflected and uninflected forms. As soon as there is a match it is assumed by the program, that it is the only possible match, homonymy thus being discarded. It can therefore happen that noun “vara” can be assumed to be a verb “vara”.

5.2.2 Towards expanding of the system

The system can be expanded in a number of ways. To start with, more exercises for lexical training can be added, provided other resources than SUC are used in the system. Using some lexicon database can facilitate generating glossaries and definition exercises; having access to some morphological database can ensure exercises for training word-building; access to the Swedish WordNet or to a synonym lexicon can make it possible to produce synonym/antonym items. Exercises on collocations need some database of collocations and idioms and a tool that will analyze collocations in texts. Still, the main source of texts should preferably be SUC.

Creating glossaries can be refined so that every word or only target words are hyperlinked to a dictionary entry (provided text words are tagged; if the text does not come from SUC PoS-tagger should be a requirement then). Hyperlinking in itself would probably not present a lot of problems. The selection of suitable concordance examples, however, is a complex question requiring deeper research. Another possible refinement is complementing the existing frequency lists with spoken frequencies for more appropriate selection of target vocabulary for glossaries.

Grammar exercises can be added. There are a number of different exercises that can be created automatically, starting from putting one lemma into different forms to exercises on agreement between nouns and verbs, exercises on tenses, comparative forms of adjectives, subjunctive mood, putting all verbs in the text into necessary tense (which would mean all analytic forms like “har blivit” should be reduced to one verb “att bli”, i.e. all auxiliary verbs should be distracted to avoid giving out unnecessary clues), etc. On syntactic level, word order exercises could be implemented and use of adverbials could be trained on automatically generated exercises.

On discourse level, provided the list of discourse markers is compiled, their use could be trained as well e.g. by providing a text with all discourse markers withdrawn from the text into a list and offering students to find appropriate place for them to ensure discourse coherence.

Reading comprehension exercises can also be added and a lot of other types of exercises. Eventually even automatic generation of tests consisting of any number and type of test items should be made possible.

On a more advanced implementation level, tools for generating feedback on learner performance and techniques for analysis and scoring of free answers can be implemented or reused. And probably many more options that I cannot think of at the moment.

5.2.3 Towards a better user interface

At present, the program lacks any user intervention into the process of creation of exercises except that the learner level and the frequency band/wordclass(es) are marked by the user. Teacher-correction window that will ensure proof-reading and disambiguation possibility is definitely a necessity. An extra window for manipulation of exercises in any way should also be an advantage.

The user interface as it looks now has not been analyzed for user-friendliness since the focus in the thesis has been made on the functionality of the system.

5.2.4 Towards improved presentation and user adaptability

The idea has been that all generated tests and exercises should be saved in QTI format and automatically sent to ITG system. QTI – Question and Test Interoperability – is a guideline for creating teaching materials for banking of teaching items issued by IMS global learning consortium.

The IMS Question & Test Interoperability (QTI) specification describes a data model for the representation of question (assessmentItem) and test (assessmentTest) data and their corresponding results reports. Therefore, the specification enables the exchange of this item, test and results data between authoring tools, item banks, test constructional tools, learning systems and assessment delivery systems. The data model is described abstractly, using [UML] to facilitate binding to a wide range of data-modelling tools and programming languages, however, for interchange between systems a binding is provided to the industry standard eXtensible Markup Language [XML] and use of this binding is strongly recommended. The IMS QTI specification has been designed to support both interoperability and innovation through the provision of well-defined extension points. These extension points can be used to wrap specialized or proprietary data in ways that allows it to be used alongside items that can be represented directly. (<http://www.imsglobal.org/question/>)

In this version of the generator QTI format has not been implemented. It is left for the future.

Apart from QTI, the issue of adaptability of a system to a student proficiency level has been left for future work. It is desirable to enrich the system with a module for creating student model of competence based on ability in different linguistic skills, e.g. vocabulary, grammar, etc.

5.2.5 Experiments and tests

Evaluation of the application can be done in several ways. Borin (2005) describes three ways: individual review, group review and formative or summative evaluation in real-life context. Each of those can describe users' conclusions of user interface, functionality of an application, economic justifiability, how pedagogical issues are dealt with in an application, and ways for further improvement.

In some future it is planned to make a pilot test of the application first comparing times that the generation of exercises/tests takes if produced manually by teacher and automatically by the application. The resulting exercises will be offered to a group of students, so that they can work with both manually and automatically produced exercises and compare them as to how clear/unclear, easy/difficult, etc. they are. Teachers will be asked to fill in an evaluation form as well.

Using LexLIX as a predictor of reading and lexical complexity of a text also needs testing. The best evaluation would be to analyze whether the texts selected for each level

are appropriately selected or not. This might need working with some groups of students of different levels during a course and collect information as the course progresses.

Another option for testing LexLIX is to collect texts used by teachers for different proficiency levels and run LexLIX analysis on them and then see how LexLIX score correspond to each proficiency level.

Selection of distractors for multiple-choice items can also become a candidate for future evaluation.

5.2.6 Other areas of application of the generator

Apart from language training purposes the generator can even be applied to other subjects as a generator of reading comprehension exercises. Multiple-choice exercises, c-tests and word bank exercises can be generated from any text, i.e. even specific in nature. Learners that have read some material sometimes need to be specifically drawn to the importance of certain passages. This can be done by way of withdrawing certain text words, preferably terminology. If a bank of specific texts is collected and, for instance, a list (or lists – if areas are different) of terminology is made, then exercises can be created for training understanding of terminology in context. An even better option would have been to make a corpus of texts with terminology mark-up included – which, of course, is a more expensive option.

5.3 Results

The objective of this research has been to create a system that, being manually fed with the level of a student, frequency band and/or wordclass(es) can automatically select text material of an appropriate level and on its basis automatically create vocabulary items. Apart from this, a number of questions have been raised and claimed to be answered by the end of the research. As expected, the output of this research can cover only part of the raised questions and claims:

- the exercise generator has demonstrated how effectively SUC can be used for the purposes of automatic generation of exercises. However, the claim to answer “what aspects of word knowledge can be trained by computer-generated exercises based on SUC and to what effect” cannot be answered on the basis of this work. The implemented authoring tool covers only a small part of all possible types of exercises. If in addition to SUC other resources can be added, most probably the major part of the aspects of word knowledge will be covered. “To what effect” also needs more research and real life experiments.
- general conclusions about which resources and technologies need to be developed to cover some of the uncovered by the generator aspects of word knowledge can be drawn, though not backed up by practical or theoretical findings. Some of the resources that are available at present are mentioned below, as well as those that are lacking but are desirable.

- general conclusions about which aspects of word knowledge cannot be trained via automatically generated exercises based on SUC are only a guess; no comprehensive analysis of this aspect has been carried out.

Table 15. Aspects of word knowledge.

<u>Form:</u>	spoken (recognition in speech, pronunciation) written (recognition in texts, spelling) word parts (morphology: inflection, derivation, word-building)
<u>Meaning:</u>	form and meaning concept and referents associations
<u>Use:</u>	grammatical functions collocations constraints on use: register/frequency/etc

It has turned out to be impossible to answer these questions to full extent. I have limited myself to studying SUC on the four types of items. These items – yes/no, wordbank, multiple-choice items and c-tests – can be effectively produced on the basis of SUC and frequency lists derived from SUC. Readability measures regulate the text choice. Word form, and its spelling, meaning of the words, grammatical functions and typical context and collocations they enter can be trained by these types of items, see Table 15. The items, however, do not cover the whole specter of word knowledge. To cover the other aspects, a number of resources are needed: explanatory lexicons, WordNets, tools for identification collocations and idioms, morphological databases, and some others. More about this has been said in subchapter “6.2 Future of the System”.

A number of tools and resources necessary for further development of the system are available, and can be reused, e.g. spell-checker STAVA; monolingual lexical database Svenska Ord - a specifically designed lexicon for learners of Swedish as a second language; synonym database “Folkets synonymlexikon”, etc. Even though available, some of those resources are not fully suitable for the exercise generator. To name a few drawbacks, Svenska Ord contains only 20,000 words, which fails to cover all necessary vocabulary for learners or for checking coined pseudowords so that they are not accidentally real words. Folkets synonymlexikon contains pairs of synonyms, which have never been proofread by specialists, but have been “voted for” by the users of Lexin webpage. Some of those pairs contain words of different wordclasses. Yet, it is worth testing them in an ICALL application before final conclusions are drawn.

Other resources that eventually can be used in the generator are WordNet that is not yet finished; morphological database that has not yet been made accessible; lemmatizers and PoS-taggers that are available for Swedish at present, yet time has not allowed us to adapt them to the needs of this generator.

Finally, certain resources do not exist at the moment of writing this thesis. Among them – tools for assessing free response items, for training productive aspects of vocabulary use, for distinguishing different meanings of a lemma (semantic disambiguation), for analysis of collocations and idioms.

Although the questions raised at the beginning are left partly unanswered, it is possible to state the impact of corpora in the field of computerized (computer-assisted) generation of

teaching materials. SUC with its annotation and markup, its selection of texts and its structure can be the central part of any ICALL system aimed at automatic production of text-based materials for different linguistic skills: reading comprehension, grammar, vocabulary, morphology. It has been demonstrated that with SUC as the sole resource it is possible to generate varied pedagogically acceptable exercises.

To my knowledge, there is no other system for Swedish that can automatically generate the same types of exercises for vocabulary training.

References

- Aist, G. (2001). Towards automatic glossarization: automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment. *International Journal of Artificial Intelligence in Education* 12, 212-231.
- Alderson, J.C. (2000). *Assessing Reading*. Cambridge University Press.
- Alderson, J.C., Clapman, C. and Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge University Press.
- Ammerlaan, T. (2002). www.WorldWideWriting.com: Developing a Multi-Lingual Process-Oriented Feedback Programme. *Annual Conference of the European Association for Computer-Assisted Language Learning (EUROCALL)* Jyvaskyla, Finland.
- Babic, S. (2002). Didax – a System for Online Testing: Technical Documentation for the Current Implementation of Teacher Client 2. p.25-35. Reports from Uppsala Learning Lab — Digital Resources in the Humanities (DRHum) project. Research reports (DRHumR).
- Bachman, L. F. (1998). Language Testing - SLA Research Interfaces. In L.F. Bachman and A.D. Cohen (eds.), *Interfaces Between Second Language Acquisition and Language Testing Research* 177-195. Cambridge University Press.
- Bachman, L. F. and Cohen, A.D. (1998). Language Testing - SLA interfaces: An Update. In L. F. Bachman and A.D. Cohen (eds.), *Interfaces Between Second Language Acquisition and Language Testing Research* 1-31. Cambridge University Press.
- Bengtsson, C. and Lingdell, M. (2002). Didax – a System for Online Testing: Technical Documentation. p.1-25. Reports from Uppsala Learning Lab — Digital Resources in the Humanities (DRHum) project. Research reports (DRHumR).
- Bergström, H. (2007). Evaluation of a Computer-Assisted Language Learning System for Swedish Language Learners. *KTH Computer Science and Communication, Centre for Speech Technology* 53 p. Stockholm: Royal Institute of Technology.
- Bick, E. (2001). The VISL System: Research and Applicative Aspects of IT-based learning. *Proceedings of NoDaLiDa 2001* 11. Uppsala, Sweden.
- Bick, E. (2005). Grammar for Fun: IT-based Grammar Learning with VISL. In: *Henriksen, Peter Juel (ed.), CALL for the Nordic Languages*. p.49-64. Copenhagen: Samfundslitteratur (Copenhagen Studies in Language).
- Bigert, J., Kann, V., Knutsson, O. and Sjöbergh, J. (2005). Grammar Checking for Swedish Second Language Learners. *Chapter in CALL for the Nordic Languages* p. 33-47. Copenhagen Studies in Language 30, Copenhagen Business School. Samfundslitteratur. .
- Borin, L. (1998). ETAP: Etablering och annotering av parallellkorpus för igenkänning av översättningsekvivalenter. *ASLA-information* 24:1, 33-40.
- Borin, L. (2002a). What Have You Done for Me Lately? The Fickle Alignment of NLP and CALL. *Presentation at the EuroCALL 2002 pre-conference workshop on NLP in CALL* Jyväskylä, Finland.
- Borin, L. (2002b). Where Will the Standards for Intelligent Computer-Assisted Language Learning Come from? *LREC 2002. Third International Conference on Language*

- Resources and Evaluation. Workshop Proceedings. International standards of terminology and language resources management.* p.61-68. Las Palmas: ELRA.
- Borin, L. (2003). The SVANTE project. <http://www.svenska.gu.se/~svelb/svante/>, accessed on October, 19, 2007.
- Borin, L., Åkerman Sarkisian, K. and Bengtsson, C. (2001). A Stitch in Time: Enhancing University Language Education with Web-based Diagnostic Testing. . *20th World Conference on Open Learning and Distance Education. The Future of Learning -Learning for the Future: Shaping the Transition.* Düsseldorf, Germany, 01-05 April 2001. .
- Borin, L. and Cerratto, T. (2002). Swedish as a Second Language and Computer Aided Language Learning. Overview of the research area. Department of Numerical Analysis and Computer Science.
- Borin, L. and Dahllöf, M. (1999). A Corpus-Based Grammar Tutor for Education in Language and Speech Technology. *In EACL'99. Computer and Internet Supported Education in Language and Speech Technology. Proceedings of a Workshop Sponsored by ELSNET and the Association for Computational Linguistics* p.36-43. Bergen, Norway. University of Bergen.
- Borin, L. and Prütz, K. (2004). New Wine in Old Skins? A Corpus Investigation of L1 Syntactic Transfer in Learner Language. In G. Aston, S. Bernardini and D. Stewart (eds.), *Corpora and Language Learners* 67-87. Amsterdam: John Benjamins.
- Borin, L. and Saxena, A. (2004). Grammar, Incorporated. In Peter Juel Henriksen (ed.), *CALL for the Nordic Languages. Copenhagen Studies in Language* 30. p.125-145. Copenhagen: Samfundslitteratur.
- Brindley, G. (1988). Describing Language Development? Rating Scales and SLA. In L. F. Bachman and A.D. Cohen (eds.), *Interfaces Between Second Language Acquisition and Language Testing Research* 112-140. Cambridge University Press.
- Brown, J.C. and Eskenazi, M. (2004). Retrieval of Authentic Documents for Reader-Specific Lexical Practice. *Proceedings of InSTIL/ICALL Symposium Venice, Italy.*
- Brown, J.C., Frishkoff, G.A. and Eskenazi, M. (2005). Automatic Question Generation for Vocabulary Assessment. *In Proceedings of HLT/EMNLP 2005.* Vancouver, B.C.
- Burstein, J., Chodorow, M. and Leacock, C. (2003). Criterion: Online Essay Evaluation: An Application for Automated Evaluation of Student Essays. *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence Acapulco, Mexico.*
- Burstein, J., Wolff, S. and Chi, L. (1999). Using Lexical Semantic Techniques to Classify Free-Responses. *Breadth and Depth of Semantic Lexicons* p. 1-18.: Kluwer Acad. Press.
- Carlberger, J., Domeij, R., Kann, V. and Knutsson, O. (2004). The Development and Performance of a Grammar Checker for Swedish: A language Engineering Perspective. *Natural Language Engineering* 1 (1).
- Carlson, L., Grönroos, M. and Lemmilä, S. (2005). Squirrel Two: Experiments on a Metasearch Engine for CALL. *NODALIDA 15* Joensuu, Finland.

- Cedergren, M. (1992). Kvantitativa läsbarhetsanalyser som metod för datorstött granskning. http://iplab.nada.kth.se/pub_all.jsp (Retrieved 2007-02-08)
Stockholm: Inst.för Numerisk analys och datalogi, Kungl. Tekniska högskolan, NADA.
- Chapelle, C.A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research* 10, 157-187.
- Chapelle, C.A. (1998). Construct Definition and Validity Inquiry in SLA Research. In Cohen A.D. Bachman L. F. (ed.), *Interfaces Between Second Language Acquisition and Language Testing Research* 32-70. Cambridge University Press.
- Cole, R. (ed) (1997). *Survey of the State of the Art in Human Language Technology*. Cambridge University Press.
- Collins-Thompson, K. and Callan, J. (2004). A Language Modelling Approach to Predicting Reading Difficulty. *Proceedings of the HLT/NAACL Annual Conference*. Boston, MA, USA.
- Collins-Thompson, K. and Callan, J. (2007). Automatic and Human Scoring of Word Definition Responses. *Proceedings of NAACL HLT 2007* 476-483. Rochester, NY.
- Coniam, D. (1997). A Preliminary Inquiry Into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests. *CALICO Journal* 14, 15-33.
- Dikli, S. (2006). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment*. Vol. 5, Nr.1, 1-36.
- Dodd, B. (1997). Exploiting a Corpus of Written German for Advanced Language Learning. In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.), *Teaching and Language Corpora* 131-145. London and New York: Longman.
- Domeij, R., Hollman, J. and Kann, V. (1994). Detection of Spelling Errors in Swedish not Using a Word List en Clair. *Journal of Quantitative Linguistics* 1, 195-201.
- Dorr, B., Hendler, J., Blanksteen, S. and Migdalof, B. (eds.) (1993). *Use of Lexical Conceptual Structure for Intelligent Tutoring*. University of Maryland, College Park, MD.
- Elert, C.-C. (1970). *Ljud och ord i svenskan*. Stockholm: Almqvist & Wiksell.
- Eriksson, E., Bälter, O., Engwall, O., Öster, A.-M. and Kjellström (formerly Sidenbladh), H. (2005). Design Recommendations for a Computer-based Speech Training System Based on End-user Interviews. In *Proceedings of International Conference on Speech and Computer*, p.483-486.
- Eskenazi, M. (1999). Using Automatic Speech Processing for Foreign Language Pronunciation Tutoring: Some Issues and a Prototype. *Language Learning and Technology* Volume 2, Number 2, p.62-76.
- Eugenio, B. Di , Fossati, D. , Yu, D. , Haller, S. and Glass, M. (2005). Natural Language Generation for Intelligent Tutoring Systems: a Case Study. . *Proceedings of the Twelfth International Conference on Artificial Intelligence in Education* Amsterdam, The Netherlands.
- Foltz, P.W., Gilliam, S. and Kendall, S. (2000). Supporting Content-Based Feedback in On-Line Writing Evaluation with LSA. *Interactive Learning Environments* Vol.8, No.2, p.111-127.

- Forsbom, E. (2006). Deriving a Base Vocabulary Pool from the Stockholm Umeå Corpus. <http://stp.lingfil.uu.se/~evafo/resources/baseformmodels/> (retrieved on December, 5, 2006).
- Fulcher, G. (1997). Text Difficulty and Accessibility: Reading Formulae and Expert Judgement. *System* vol.25, 497-513.
- Gardner, D. (2007). Validating the Construct of Word in Applied Corpus-based Vocabulary Research: A Critical Survey. *Applied Linguistics* 28/2, p.241-265.
- Gavioli, L. (1997). Exploring Texts through the Concordancer: Guiding the Learner. In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.), *Teaching and Language Corpora* 83-99. London and New York: Longman.
- Godfrey, J.J. and Zampolli, A. (1997). Language Resources. Overview. In R. Cole (ed.), *Survey of the State of the Art in Human Language Technology* p.381-384. Cambridge University Press and Giardini Editori e Stampatori in Pisa.
- Graesser, A.C. and Wisher, R.A. (2001). Question Generation as a Learning Multiplier in Distributed Learning Environments. Army research inst for the behavioral and social sciences Alexandria VA. Report Number A654993, 2001.
- Greenfield, J. (2004). Readability Formulas for EFL. *JALT Journal*.
- Gyllstad, H. (2004). Testing L2 Vocabulary: Current Test Formats in English as a L2 Used at Swedish Universities. *The department of English in Lund: Working Papers in Linguistics* 4, 21-40.
- Haller, S. and Eugenio, B. Di (2003). Minimal Text Structuring to Improve the Generation of Feedback in Intelligent Tutoring Systems. *Proceedings of the Sixteenth Florida Artificial Intelligence Research Conference* 382-386. St. Augustine, FL.
- Hammarberg, B. (2005). *Introduktion till ASU-korpusen, en longitudinell muntlig och skriftlig textkorpus av vuxna inlärares svenska med en motsvarande del från infödda svenskar*. Institutionen för lingvistik, Stockholms Universitet.
- Hammarström, H. (2002). Overview of IT-based Tools for Learning and Training Grammar. *Project report, IT-based Collaborative Learning in Grammar*: Department of Linguistics, Uppsala University.
- Hassel, M. (2001). Internet as Corpus - Automatic Construction of a Swedish News Corpus. *In the Proceedings of NODALIDA'01 - 13th Nordic Conference on Computational Linguistics*. Uppsala, Sweden.
- Heilman, M., Collins-Thompson, K., Callan, J. and Eskenazi, M. (2006). Classroom Success of an Intelligent Tutoring System for Lexical Practice and Reading Comprehension. *ICSLP*.
- Heilman, M., Collins-Thompson, K., Callan, J. and Eskenazi, M. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. *Proceedings of NAACL HLT 2007* 460-467. Rochester, NY.
- Heilman, M. and Eskenazi, M. (2006). Language Learning: Challenges for Intelligent Tutoring Systems. *Workshop on Ill-defined Domains in Intelligent Tutoring*, Taiwan: http://reap.cs.cmu.edu/Papers/ITS06_illdefinedworkshop_HeilmanEskenazi.pdf.
- Higgins, J. (1995). *Computers and English Language Learning*. Oxford, England: Intellect Ltd.

- Higgins, J. and Johns, T. (1984). *Computers in Language Learning*. London and Glasgo: Collins ELT.
- Hjalmarsson, A., Wik, P. and Brusik, J. (2007). Computer Assisted Conversation Training for Second Language Learners. *In Fonetik 2007* Stockholm, Sweden.
- Huitbregtse, I., Admiraal, W. and Meara, P. (2002). Scores on a Yes-No Vocabulary Test: Correction for Guessing and Response Style. *Language Testing* 19, 227-245.
- Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge, UK: Cambridge University Press.
- Johansson, P. (2006). Klustring av svenska texter. Magisteruppsats. *KTH: NADA* 20 p. Stockholm: KTH.
- Johnson, W., Vilhjalmsen, H. and Marsella, S. (2005). Serious Games for Language Learning: How Much Game, How Much AI? *In 12: th International Conference on Artificial Intelligence in Education*. Amsterdam.
- Jung Hee, K., Freedman, R., Glass, M. and Evens, M. W. (2006). Annotation of Tutorial Dialogue Goals for Natural Language Generation. *Discourse Processes* vol. 42, pp.37-74.
- Källgren, G., Gustafson-Capková, S. and Hartmann, B. (2006). Manual of the Stockholm Umeå Corpus version 2.0. 85 p.
- Kann, V. (2003). Lägesrapport för CrossCheck för perioden 1 juli-31 december 2003. <http://www.csc.kth.se/tcs/projects/xcheck/lagesrapport-031231.html>, accessed 2007-11-06 Stockholm: KTH: NADA.
- Kempen, G. (1996). Human Language Technology Can Modernize Writing and Grammar Instruction. *COLING - 96. The 16th international conference on computational linguistics. Proceedings, vol.2*. 1005-1006. Copenhagen: Center for Språkteknologi.
- Kintsch, E., Steinhart, D., Stahl, G. and Group, LSA Research (2000). Developing Summarization Skills through the Use of LSA-Based Feedback. *Interactive Learning Environments* Vol.8, No.2, p.87-109.
- Knutsson, O. (2005). Developing and Evaluating Language Tools for Writers and Learners of Swedish. Doctoral Thesis. *KTH Computer Science and Communication* Stockholm: KTH.
- Knutsson, O., Cerratto Pargman, T., Severinson Eklundh, K. and Westlund, S. (2005). Designing and Developing a Language Environment for Second Language Writers. *Computers and Education. Elsevier*.
- Kotsinas, U.-B. (2005). *Invandrarsvenska*. Uppsala: Hallgren & Fallgren.
- Kukich, K. (1992). Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys* 24, 377-439.
- Kunichika, H., Minoru, H., Tsukasa, H. and Akira, T. (2003). A Method to Resolve Ambiguity of Interpretation of English Sentences for Intelligent English Learning Support Systems. <http://www.isl.hiroshima-u.ac.jp/html/tsukasa/assets/pdf-files/Kunichika/ICCE2003.pdf> (Retrieved 2006-05-28).
- Kunichika, H., Minoru, U., Tsukasa, H. and Akira, T. (2005). Realizing Adaptive Questions and Answers for ICALL Systems. <http://www.isl.hiroshima-u.ac.jp/html/tsukasa/assets/pdf-files/Kunichika/AIED2005.pdf> (Retrieved 2006-05-27).

- Laufer, B. and Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics* 16, 307-322.
- Laurillard, D. (2002). *Rethinking University Teaching: A Framework for the Effective use of Educational Technology*. London: Routledge.
- Leech, G. (1997). Teaching and Language Corpora: a Convergence. In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.), *Teaching and Language Corpora*. London and New York: Longman.
- Li, T. and Sambasivam, S. (2005). Automatically Generating Questions in Multiple Variables for Intelligent Tutoring. *Issues in Informing Science and Information Technology*.
- Lindberg, J. and Eriksson, E. (2005). CrossCheck-korpusen - en elektronisk svensk inlärnarkorpus. <http://www.csc.kth.se/tcs/projects/xcheck/korpus.html>, accessed 2007-11-06 Stockholm: KTH: NADA.
- Lindberg, J. and Eriksson, G. (2004). CrossCheck-korpusen - en elektronisk svensk inlärnarkorpus. *Proceedings of the ASLA Conference 2004 Södertörns högskola, Sweden*.
- Lu, X. (2006). Expert Tutoring and Natural Language Feedback in Intelligent Tutoring Systems. *Doctoral Student Consortium at the 14th International Conference on Computers in Education (ICCE2006)* Beijing, China.
- Ma, Q. and Kelly, P. (2006). Computer-Assisted Vocabulary Learning: Design and Evaluation. *Computer-Assisted Language Learning* 19, 15-45.
- McEnery, T. and Wilson, A. (2001). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Meara, P. (2005). Lexical Frequency Plorifles: A Monte Carlo Analysis. *Applied Linguistics* vol.26/1, p.32-46.
- Meskill, C., Anthony, N., Hilliker-Vanstrander, S., Tseng, C. and You, J. (2006). CALL: A Survey of K-12 ESOL Teacher Uses and Preferences. *TESOL Quarterly* 40, 439-451.
- Meyer, C. (2002). *English Corpus Linguistics*. Cambridge: Cambridge University Press.
- Mindt, D. (1997). Corpora and the Teaching of English in Germany. In Fligelstone S. Wichmann A., McEnery T., Knowles G. (ed.), *Teaching and Language Corpora* 40-50. New York: Addison Wesley Longman Inc.
- Minugh, D. (1997). All the Language that's Fit to Print: Using British and American Newspaper CD-ROMs as Corpora. In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.), *Teaching and Language Corpora* 67-82. London and New York: Longman.
- Mitkov, R. and Ha, L.A. (2003). Computer-Aided Generation of Multiple-Choice Tests. *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing*, 17-22.
- Monaghan, W. and Bridgeman, B. (2005). E-Rater as a Quality Control on Human Scores. *ETS R&D Connections*: Princeton, NJ: ETS.
- Nagata, N. (1997). An Experimental Comparison of Deductive and Inductive Feedback Generated by a Simple Parser. *System* Vol.25, No.4, pp.515-534.
- Nation, P. and Waring, R. (1997). Vocabulary Size, Text Coverage and Word Lists. *Vocabulary: Description, Acquisition and Pedagogy*, 6-19.

- Nerbonne, J. and Smit, P. (1996). GLOSSER-RuG: In Support of Reading. *COLING-96. The 16th International Conference on Computational Linguistics. Proceedings, vol.2* 830-835. Copenhagen: Centre for Sprogteknologi.
- Nilsson, K. (2003). A Meta Search Approach to Locating and Classifying Reading Material for Learners of Nordic Languages. *Department of Linguistics* 44 p. Uppsala: Uppsala University.
- Nilsson, K. and Borin, L. (2002). Living off the Land: The Web as a Source of Practice Texts for Learners of Less Prevalent Languages. *Proceedings of LREC 2002, Third International Conference on Language Resources and Evaluation* p.411-418. Las Palmas: ELRA.
- Nivre, J., Nilsson, J. and Hall, J. (2006). Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. . *In Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)* Genoa, Italy.
- Niwinski, W. (2002). CALLe svenska: an Infinity of Exercises, a Single Student in Focus. *Computer Assisted Language Learning* Volume 15, Issue 1, p.85-90.
- O'Keefe, A. and Farr, F. (2003). Using Language Corpora in Initial Teacher Education: Pedagogic Issues and Practical Applications. *TESOL Quarterly* 37, 389-418.
- O'Keefe, A., McCarthy, M. and Carter, R. (2007). *From Corpus to Classroom. Language Use and Language Teaching*. Cambridge University Press.
- Olausson Källfelt, A. and Fogelberg, M. (2004). Utvärdering av Lingus - ett system för datorstödd språkkärning. *Datalogivstprogrammet, Institutionen för lingvistik* 43 p. Gothenburg: Gothenburg university.
- Oller, J.W., Jr. (1973). Cloze Tests of Second Language Proficiency and What They Measure. *Language Learning* 23, 105-118.
- Olsson, L.-J. and Borin, L. (2000). A Web-Based Tool for Exploring Translation Equivalents on Word and Sentence Level in Multilingual Parallel Corpora. *In Erikoiskielet ja kännösteoria - Fackspråk och översättningsteori - LSP and Theory of Translation. 20th VAKKI Symposium*, p.76-84. Vaasa, Finland, University of Vaasa.
- Peters, P. (1997). Micro- and Macrolinguistics for Natural Language Processing. In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.), *Teaching and Language Corpora* p.175-185. London & New York: Longman.
- Ramsden, P. (2002). *Learning to Teach in Higher Education*. London and New York: Routledge.
- Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press.
- Riedel, E., Dexter, S., Scharber, C. and Doering, A. (2005). Experimental Evidence on the Effectiveness of Automated Essay Scoring in Teacher Education Cases. *86th Annual Meeting of the American Educational Research Association* Montreal, CA.
- Saxena, A. and Borin, L. (2002). Locating and Reusing Sundry NLP Flotsam in an e-Learning Application. *LREC 2002. Workshop Proceedings. Customizing Knowledge in NLP Applications: Strategies, Issues and Evaluation* 45-51. Las Palmas, Spain.
- Shohamy, E. (1998). How Can Language Testing and SLA Benefit from Each Other? The case of Discourse. In L. F. Bachman and A.D. Cohen (eds.), *Interfaces*

- Between Second Language Acquisition and Language Testing Research* 156-176. Cambridge University Press.
- Sigurd, B. (1965). *Phonotactic Structures in Swedish*. Akademisk avhandling: Lund, Berlingska Boktryckeriet.
- Svensson, J. (1999). Med fokus på dagens språk. Professoninstallation 8/10-99. http://www3.lu.se/info/profinst/9910/11_svensson.html, accessed 2007-11-07: Lunds Universitet.
- Tufis, D. (1996). CALL: The Potential of Lingware and the Use of Empirical Linguistic Data. *COLING - 96. The 16th international conference on computational linguistics. Proceedings, vol.2.* 1010-1011. Copenhagen: Center for Språkteknologi.
- Valenti, S., Neri, F. and Cucchiarelli, A. (2003). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education* vol.2, p.319-330.
- Wik, P. (2004). Designing a Virtual Language Tutor. . In *Proc of The XVIIth Swedish Phonetics Conference, Fonetik 2004.* p. 136-139. Stockholm University, Sweden.
- Williams, R. and Dreher, H. (2005). Telecommunications Use in Education to Provide Interactive Visual Feedback on Automatically Graded Essays. In R. Cooper, G. Madden and A. Lloyd (eds.), *International Telecommunications Society Asia-African Australasian Conference* Perth, WA.
- Wilson, E. (1997). The Automatic generation of CALL Exercises from General Corpora. In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.), *Teaching and Language Corpora* London and New York: Longman.
- Zareva, A. (2005). Models of Lexical Knowledge Assessment of Second Language Learners of English at Higher Levels of Language Proficiency. *System* 33, 547-562.
- Zock, M. (1996). Computational Linguistics and its Use in Real World: the Case of Computer-Assisted Language Learning. *COLING - 96. The 16th international conference on computational linguistics. Proceedings, vol.2.* 1001-1004. Copenhagen: Center for Språkteknologi.

Appendices.

Appendix 1. Corpora of Swedish

Corpora of Written Swedish (non-commercial)

- Konkordanser – a collection of corpora with integrated concordance tool, not annotated. Available corpora are listed below. Source: <http://spraakbanken.gu.se/>
 - Press 65
 - Press 76
 - DN 1987
 - Press 95
 - Press 96
 - Press 97
 - Press 98
 - GP 01
 - GP 02
 - GP 03
 - GP 04
 - F&F (Forskning och Framsteg) (= magazine "Science and Progress")
 - Äldre svenska romaner (= older Swedish novels)
 - 1800-tals romaner (= 19th century novels)
 - B.romaner I (= novels1)
 - B.romaner II (= novels2)
 - Strindberg brev (= Strindberg's letters)
 - Strindberg romaner (= Strindberg's novels)
 - SAOL 11 (Svenska Akademiens Ordlista)
 - Psalmboken (= Book of Psalms)
 - Sv.förf.samling (= collection of Swedish writers)
- ORDAT Svenska dagbladets årsbok 1923 – 1958 - a collection of newspaper articles (<http://spraakbanken.gu.se/>)
- SNP – preliminary reports of the proceedings from the Swedish Parliament 78-79 (<http://spraakbanken.gu.se/>)
- PAROLE is a POS-tagged and morphologically annotated corpus of several European languages, including Swedish. Swedish part comprises approximately 19,4 mln. words coming from novels, newspapers, magazines and other sources. Texts date back to 1976-1997 (<http://spraakbanken.gu.se/>)
- Bellman - C.M. Bellman's literary works (<http://spraakbanken.gu.se/>)
- Strindberg - Strindberg's collection (<http://spraakbanken.gu.se/>)
- Talbanken (in MAMBA version and newer annotation versions) is a treebank, consisting of both written and spoken parts, the written part containing a professional native speaker part and a learner part represented by upper secondary pupils with Swedish as their mother tongue (Nivre, Nilsson & Hall 2006). The corpus is POS-tagged and syntactically annotated.
It is available from <<http://w3.msi.vxu.se/~nivre/research/Talbanken05.html>>
- Skrivsyntax seems to be a part of Talbanken, since Talbanken is a result of two projects – Skrivsyntax and Talsyntax (Svensson 1999). Available in ITG system
- ASU (Andraspråkets StrukturUtveckling) – written part – consists of essays written by native speakers and learners, POS-tagged (available in ITG).
- SUC (Stockholm Umeå Corpus) (<http://spraakbanken.gu.se/>), POS-tagged.
- Litteraturbanken – a collection of computerized versions of classic Swedish literature. It contains both older texts and modern texts, and has a concordance software. Litteraturbanken is publicly available and free of charge. Available at

<http://litteraturbanken.se/>

- Project Runeberg – online library, rather than corpus... (<http://runeberg.org/>)
- eBoklagret – online library (<http://www.omnibus.se/eBoklagret/>)
- PressText – commercial archive (<http://www.presstext.se/>) & mediaArkivet (<http://www.retriever-info.com/>)
- Swedish novels 80-81 - 3.7 million words (<http://www.ling.gu.se/projekt/tal/>)
- Läkemedelsboken (medicine book) - 380 000 words (<http://www.ling.gu.se/projekt/tal/>)
- Fass (medicine) - 1 million words (<http://www.ling.gu.se/projekt/tal/>)
- VMDM (medicine) - 590000 words (<http://www.ling.gu.se/projekt/tal/>)
- The Bible (1917) - 800000 words (<http://www.ling.gu.se/projekt/tal/>)
- GöteborgsPosten 1993-2001 (newspaper) - 795 000 articles, 190 million words. (<http://www.ling.gu.se/projekt/tal/>)
- Helsingborgs Dagblad 1994-2001 (newspaper) - 570 000 articles, 140 million words. (<http://www.ling.gu.se/projekt/tal/>)
- Norrköpings Tidningar, Nya Dagen, NorrlandsKuriren (newspaper) - 130 000 articles, 60 million words. (<http://www.ling.gu.se/projekt/tal/>)
- SynTag (Nivre et al. 2006)
- ETAP - is the acronym of the project title “Etablering och annotering av parallellkorpus för igenkänning av översättningsekvivalenter” (Olsson & Borin 2000). ETAP is an annotated parallel corpus consisting of three main parts (Borin 1998):
 1. technical documents in Swedish, English, Finnish, French, Italian, Dutch, Spanish and German
 2. Invandrartidningen – several issues of the magazine for immigrants in Swedish, English, Finnish, Polish, Serbo-Croatian and Spanish
 3. Regeringsförklaring from Swedish government from 1998 and on in Swedish, English, French, Spanish and German
- OrdiL (Ordförråd i läromedel). It is an ongoing project in Swedish as L2 where texts used in course books in Swedish compulsory school in Science (Biology, Chemistry, Physics), Society-oriented subjects (Geography, History, Psychology, Religion and Social Studies) and Mathematics are collected into different subcorpora with the aim to analyze vocabulary frequencies and find out core vocabulary that non-Swedish pupils should get help with in the first place. Whether the corpora will be made available is unclear
- KTH News Corpus (Hassel 2001; Johansson 2006) is an automatically constructed corpus from news texts available on Internet. Texts are collected, clustered acc.to topics; in 2001 there was a plan to automatically tag words and lemmatize them. In 2001 the copyright issues were not resolved and corpus could be used only for academic research within NADA's research group.
- Karolinska Institutets medicinska textsamling (Johansson 2006)

Corpora of Spoken Swedish

- Talbanken (in MAMBA version and newer annotation versions) is a treebank, consisting of both written and spoken parts, the written part containing a professional native speaker part and a learner part represented by upper secondary pupils with Swedish as their mother tongue (Nivre et al. 2006). The corpus is POS-tagged and syntactically annotated. It is available from <<http://w3.msi.vxu.se/~nivre/research/Talbanken05.html>>
- ASU (Andraspråkets StrukturUtveckling) – spoken part – consists of interviews with native speakers and learners. Interviews are transcribed and tagged. Available in ITG system
- Göteborg Spoken Language Corpus is an incrementally growing corpus of spoken native Swedish from different social activities, approx. 1.5 million words. Transcribed and POS-annotated. Available at <http://www.ling.gu.se/projekt/tal/>

- Adult language learners of Swedish Available at <http://www.ling.gu.se/projekt/tal/>; it is a part of EALA/ESFSD corpus
- Child language corpus (Swedish and Scandinavian), 0.5 million words including the adults Available at <http://www.ling.gu.se/projekt/tal/>
- Aphasic, deaf and dyslexic speakers. Available at <http://www.ling.gu.se/projekt/tal/>
- Child (3-6 years old) language corpus, 94 children, 260 000 words, Lisbeth Hedelin's material. Available at <http://www.ling.gu.se/projekt/tal/>
- Hedelin's material. Available at <http://www.ling.gu.se/projekt/tal/>
- WOZ Corpus, Bionic. Available at <http://www.ling.gu.se/projekt/tal/>
- Educational progress - 416 interviews, 2 million words, Kjell Härnqvist's material' Available at <http://www.ling.gu.se/projekt/tal/>

Learner Corpora

Corpora and descriptions

	Writ- ten	Spo- ken
▪ Talbanken (in MAMBA version and newer annotation versions) is a treebank, consisting of both written and spoken parts, the written part containing a professional native speaker part and a learner part, though not L2 learner (Nivre et al. 2006). The corpus is POS-tagged and syntactically annotated. It is available from < http://w3.msi.vxu.se/~nivre/research/Talbanken05.html >	X	X
▪ CrossCheck & SVANTE	X	

Within this project Cross-Check there has been collected and annotated a corpus of written learner texts under the name of Cross-Check corpus (Lindberg & Eriksson 2005). At the same time one can run into the name SVANTE (SVenska ANdraspråks TExter) which is a sub-project of CrossCheck (Borin 2003); at the same time one can see equation mark between CrossCheck Learner Corpus and SVANTE (Bigert et al. 2005). Since CrossCheck corpus contains texts produced by both native (upper-secondary school pupils) and non-native learners of Swedish, SVANTE is a unifying name for the part of CrossCheck consisting of texts produced by non-native learners of Swedish, which means SSM-part, SFI-part and Granska-part.

CrossCheck consists of the following parts:

1. SSM-part (= Svenska som Målspråk) consists of essays written during 1972-1975 by grown-up learners of Swedish with 10 different mother tongues (approx.112.000 words) (Lindberg & Eriksson 2004; Lindberg & Eriksson 2005). This part of the corpus is available through ITG system (IT-based Collaborative Learning in Grammar).
2. SFI-part consists of essays written in 1997 by 275 grown-up learners (75.000 words) of SFI (Swedish for Immigrants). Essays have been collected by Inger Lindberg, scanned and tagged by participants of CrossCheck project. This part is also available through ITG system.
3. Granska-part contains 65 text files (approximately 35.000 words) written by 10 learners of Swedish, some of the texts are rather advanced, e.g. protocols and essays. The material has been collected at Stockholm University.
4. Argus-part is a database of 287 argumentative texts (approx. 120,000 words) written by native speakers of Swedish (school pupils), located in Uppsala University.
5. In SvSFi-part 10 native speakers of Swedish were asked to write essays on the same topics as learners of Swedish. This part is used as a reference.

Some information on CrossCheck annotation, search tools and some other details about the corpus can be found in (Kann 2003)

- ASU corpus (in ITG). ASU consists of both written part (essays) and spoken part (interviews), where both native speakers and learners of Swedish are represented. POS-tagged. X X

- EALA/ESFSLD – this is a European Science Foundation Second Language Databank (ESFSLD), where spoken learner language in several European languages is collected, among them Swedish (Knutsson 2005). The information about the corpus can be can be obtained from the Institution for Linguistics, GU as well as at the following link:
http://www.mpi.nl/ISLE/overview/Overview_ESFSLD.html X

- TISUS – an ongoing project. A number of learner essays from Swedish examination TISUS have been collected and are planned to be POS-tagged and converted into XML-format. Will be added to Cross-check corpus and made available in ITG system. X

- Educational progress - 416 interviews, 2 million words, Kjell Härnqvist's material' Available at <http://www.ling.gu.se/projekt/tal/> X

Appendix 2. Function Words in 8 Frequency Bands

FB1 – Function Words

3	och.CC	32995.649859	9	och.CCS.33058	o.CCS.3
4	i.S	28081.060766	9	i.SPS.28254	
5	en.DI	25958.046833	9	ett.DI@NS@S.7952	en.DI@US@S.18050
7	den.DF	19927.775415	9	de.DF@0P@S.5934	
	dom.DF@0P@S.44			di.DF@0P@S.1	det.DF@NS@S.4499
	thet.DF@NS@S.2			den.DF@US@S.9708	
8	på.S	14763.972534	9	på.SPS.14839	
9	det.PF	14613.851763	9	det.PF@NS0@S.14925	
	de.PF@NS0@S.4			'at.PF@NS0@S.1	're.PF@NS0@S.1
	't.PF@NS0@S.1			Thet.PF@NS0@S.1	de'.PF@NS0@S.1
	dett.PF@NS0@S.1			dä.PF@NS0@S.1	et.PF@NS0@S.1
	re.PF@NS0@S.1				
10	av.S	14424.590536	9	av.SPS.14873	af.SPS.44
12	att.CI	12855.904481	9	att.CIS.12902	at.CIS.3
	å.CIS.2				
13	som.PH	12657.846815	9	som.PH@000@S.12752	
14	för.S	11647.723286	9	för.SPS.11863	[F]ör.SPS.1
15	att.CS	11526.639942	9	att.CSS.11609	at.CSS.3
16	med.S	11514.890979	9	med.SPS.11525	me.SPS.1
	mö.SPS.1				
17	till.S	9757.613351	9	till.SPS.9822	til.SPS.7
	te.SPS.1			tä'.SPS.1	
20	han.PF	7608.237099	9	honom.PF@USO@S.1398	
	han.PF@USO@S.3			han.PF@USS@S.8358	
22	jag.PF	6884.720937	9	mig.PF@USO@S.1400	mej.PF@USO@S.32
	mi.PF@USO@S.5			me'.PF@USO@S.1	jag.PF@USS@S.6729
	ja.PF@USS@S.1			ja'.PF@USS@S.1	jak.PF@USS@S.1
24	som.CC	5622.637513	9	som.CCS.5683	
25	sig.PF	5515.953389	9	sig.PF@000@S.5782	sej.PF@000@S.28
27	de.PF	5201.070850	9	dom.PF@0P0@S.211	dem.PF@0P0@S.1307
	them.PF@0P0@S.2			de.PF@0PS@S.3864	the.PF@0PS@S.2
	dem.PF@0PS@S.1				
28	men.CC	5094.237357	9	men.CCS.5282	
29	om.S	5071.926601	9	om.SPS.5125	öm.SPS.1
30	vi.PF	4720.333813	9	oss.PF@UPO@S.788	vi.PF@UPS@S.4017
32	man.PI	4432.274725	9	man.PI@USS@S.4576	
33	sin.PS	4275.134836	9	sina.PS@0P0@S.1135	
	sine.PS@0P0@S.2			sitt.PS@NS0@S.957	sin.PS@US0@S.2243
34	från.S	3995.292786	9	från.SPS.4023	frå.SPS.1
35	eller.CC	3771.504491	9	eller.CCS.3941	älr.CCS.1
38	hon.PF	3261.421389	9	henne.PF@USO@S.817	
	hon.PF@USS@S.3905				
40	om.CS	3065.843561	9	om.CSS.3111	
42	den.PF	2967.713327	9	den.PF@US0@S.3046	'n.PF@US0@S.1
49	denna.DF	2648.670499	9	dessa.DF@0P@S.888	denne.DF@MS@S.48
	detta.DF@NS@S.696			dette.DF@NS@S.1	denna.DF@US@S.1217
50	någon.DI	2632.111357	9	några.DI@0P@S.1013	
	nåra.DI@0P@S.3			något.DI@NS@S.629	nåt.DI@NS@S.28
	nåhanna.DI@NS@S.1			någon.DI@US@S.953	nån.DI@US@S.42
52	vid.S	2602.514654	9	vid.SPS.2678	wid.SPS.3
54	under.S	2531.366822	9	under.SPS.2619	

64 vad.PH 1877.339306 9 vad-.PH@NS0@C.1 vad.PH@NS0@S.1913
va.PH@NS0@S.6 hvad.PH@NS0@S.3 Hwad.PH@NS0@S.1

66 mot.S 1750.805388 9 mot.SPS.1797

67 efter.S 1699.925245 9 efter.SPS.1710

68 du.PF 1693.190449 9 dig.PF@US0@S.419 dej.PF@US0@S.35
du.PF@USS@S.1803

71 genom.S 1607.868137 9 genom.SPS.1648

72 än.CC 1590.502950 9 än.CCS.1607

73 all.D0 1550.106349 9 alla.D0@0P@S.1174 all.D0@0P@S.1
allt.D0@NS@S.197 all.D0@US@S.191

80 över.S 1372.281788 9 över.SPS.1414 öfver.SPS.2
öfwer.SPS.1

84 mellan.S 1321.937401 9 mellan.SPS.1381 mällan.SPS.1

86 någon.PI 1265.672634 9 några.PI@0P0@S.147
något.PI@NS0@S.702 nåt.PI@NS0@S.29 någon.PI@US0@S.420
nån.PI@US0@S.26

89 detta.PF 1225.691254 9 detta.PF@NS0@S.1294

93 hans.PS 1155.591684 9 hans.PS@000@S.1324

94 all.PI 1134.621929 9 alla.PI@0P0@S.464 allom.PI@0P0@S.2
allo.PI@0P0@S.1 allt.PI@NS0@S.693

95 inom.S 1109.566288 9 inom.SPS.1318

104 min.PS 1036.883261 9 mina.PS@0P0@S.204 mitt.PS@NS0@S.271
min.PS@US0@S.716

106 vår.PS 1009.628403 9 våra.PS@0P0@S.365 vårt.PS@NS0@S.243
vår.PS@US0@S.444 våran.PS@US0@S.1

113 ingen.DI 946.881593 9 inga.DI@0P@S.265 inget.DI@NS@S.215
ingen.DI@US@S.506

114 vilken.PH 923.399709 9 vilka.PH@0P0@S.220
hvilka.PH@0P0@S.2 hwilka.PH@0P0@S.1 vilket.PH@NS0@S.614
hkt.PH@NS0@S.2 hvilket.PH@NS0@S.2 vilken.PH@US0@S.148
hvilken.PH@US0@S.2 hwilken.PH@US0@S.1

121 samma.DI 854.479429 9 samma.DI@00@S.861 samme.DI@MS@S.4

124 utan.S 841.948460 9 utan.SPS.856

135 hos.S 749.109989 9 hos.SPS.775

147 enligt.S 684.986063 9 enligt.SPS.844

148 varje.DI 684.900751 9 varje.DI@0S@S.694

149 annan.PI 683.880046 9 andra.PI@0P0@S.307
androm.PI@0P0@S.1 annat.PI@NS0@S.378

151 ur.S 661.716299 9 ur.SPS.675

152 både.CC 660.395781 9 både.CCS.680

167 en.PI 606.653923 9 ett.PI@NS0@S.147 en.PI@US0@S.467
enom.PI@US0@S.1

172 bland.S 581.216280 9 bland.SPS.598

176 åt.S 559.595347 9 åt.SPS.570 åt'.SPS.2

179 deras.PS 554.843599 9 deras.PS@000@S.567

182 utan.CC 535.725603 9 utan.CCS.567

183 eftersom.CS 530.426830 9 eftersom.CSS.552

185 vilken.DH 514.739310 9 vilka.DH@0P@S.214
vilket.DH@NS@S.100 vilken.DH@US@S.213

204 hennes.PS 465.633513 9 hennes.PS@000@S.653

216 medan.CS 448.705726 9 medan.CSS.465

228 så.CC 424.840830 9 så.CCS.463

229 ingen.PI 423.661021 9 inga.PI@0P0@S.8
inget.PI@NS0@S.148 ingen.PI@US0@S.311

231 ni.PF 418.514475 9 er.PF@UPO@S.84 eder.PF@UPO@S.1
ni.PF@UPS@S.234 er.PF@US0@S.66 ni.PF@USS@S.320

241 samt.CC 397.582902 9 samt.CCS.559

245	innan.CS	393.022499	9	innan.CSS.409
255	trots.S	378.305873	9	trots.SPS.389
259	dess.PS	364.782842	9	dess.PS@000@S.392
274	framför.S	349.036026	9	framför.SPS.356
282	varandra.PF	335.802232	9	varandra.PF@0PO@S.323
	varann.PF@0PO@S.27			hvarandra.PF@0PO@S.2
320	mången.PI	290.662362	9	många.PI@0P0@S.301
	månge.PI@0P0@S.1	mångt.PI@NS0@S.2		mången.PI@US0@S.1
323	inför.S	286.159699	9	inför.SPS.296
333	kring.S	276.671944	9	kring.SPS.281
344	sedan.S	267.404754	9	sedan.SPS.281
347	före.S	266.308074	9	före.SPS.274
357	bakom.S	263.273759	9	bakom.SPS.288
360	vem.PH	260.344068	9	vem.PH@US0@S.273
				hvem.PH@US0@S.2
375	utanför.S	247.825063	9	utanför.SPS.251
381	dels.CC	243.118344	9	dels.CCS.294
383	din.PS	242.066652	8	dina.PS@0P0@S.58
	din.PS@US0@S.166			ditt.PS@NS0@S.71
389	så.CS	237.927309	9	så.CSS.246
420	dessa.PF	215.976316	9	dessa.PF@0P0@S.262
477	per.S	190.341365	9	per.SPS.228
482	ingenting.PI	189.288488	9	ingenting.PI@NS0@S.240
506	mycket.PI	178.063443	9	mycket.PI@NS0@S.189
507	liksom.S	177.992554	9	liksom.SPS.190
560	någonting.PI	158.420026	9	någonting.PI@NS0@S.148
	nånting.PI@NS0@S.35			
577	denna.PF	154.527416	9	denne.PF@MS0@S.64
	denna.PF@US0@S.101			
591	via.S	150.746712	9	via.SPS.170
593	såväl.CC	150.597279	9	såväl.CCS.196
612	runt.S	146.036252	9	runt.SPS.161
643	vars.PE	140.928145	9	vars.PE@000@S.144
				hvars.PE@000@S.1
673	sedan.CS	135.402023	9	sedan.CSS.142
693	förutom.S	132.235486	9	förutom.SPS.143
731	omkring.S	124.846095	9	omkring.SPS.130
755	för.CC	121.533994	9	för.CCS.153
818	varken.CC	109.699689	9	varken.CCS.114
				hvarken.CCS.1
822	sådan.PI	109.022123	9	sådana.PI@0P0@S.68
	såna.PI@0P0@S.6	sånt.PI@NS0@S.39		
835	å.S	107.512435	9	å.SPS.118

FB2 – Function Words

1022	var.PI	84.801103	9	var.PI@US0@S.83	hvar.PI@US0@S.2
	hvar.PI@US0@S.2	hvarjom.PI@US0@S.1			
1027	tills.CS	84.370456	9	tills.CSS.101	
1093	båda.PF	79.255948	8	båda.PF@0P0@S.87	
1106	antingen.CC	78.517864	9	antingen.CCS.91	
1127	utifrån.S	76.726628	9	utifrån.SPS.111	
1151	detsamma.PF	74.754671	9	detsamma.PF@NS0@S.79	
1161	flera.PI	73.713017	9	flera.PI@0P0@S.76	
1195	fast.CC	71.351051	9	fast.CCS.91	
1248	utom.S	67.657259	9	utom.SPS.70	
1258	ifrån.S	66.661429	9	ifrån.SPS.76	
1278	var.DI	65.190020	9	vart.DI@NS@S.15	hvert.DI@NS@S.1
	var.DI@US@S.56				
1293	intill.S	64.467175	8	intill.SPS.73	

1294	ju.CC	64.413322	9	ju.CCS.74
1322	emot.S	61.890246	9	emot.SPS.71
1390	förrän.CS	58.952411	9	förrän.CSS.65
1474	vare.CC	53.307605	9	vare.CCS.56
1523	ty.CC	50.846995	8	ty.CCS.63
1573	än.CS	48.239364	9	än.CSS.53
1577	nära.S	48.156737	9	nära.SPS.53
1581	fast.CS	47.990296	7	fast.CSS.74
1582	tills.S	47.973289	9	tills.SPS.49
1620	bredvid.S	46.595077	7	bredvid.SPS.67
1625	utöver.S	46.401490	9	utöver.SPS.60
1724	ovanför.S	43.294853	8	ovanför.SPS.53
1810	gentemot.S	40.611996	9	gentemot.SPS.52
1823	fler.PI	40.145746	9	fler.PI@0P0@S.46
1831	er.PS	39.996307	5	era.PS@0P0@S.25 ert.PS@NS0@S.16
	er.PS@US0@S.60			eder.PS@US0@S.1
1882	såsom.CC	38.338808	6	såsom.CCS.69
1896	allting.PI	38.093898	7	allting.PI@NS0@S.55
1903	ibland.S	37.928915	9	ibland.SPS.41

FB3 – Function Words

2093	dock.CC	32.618984	9	dock.CCS.39
2217	inklusive.S	30.039844	8	inklusive.SPS.35
2223	emellan.S	29.958467	7	emellan.SPS.36
2314	ovan.S	28.002388	8	ovan.SPS.35
2365	huruvida.CS	27.042156	9	huruvida.CSS.33
2427	fastän.CS	26.175925	6	fastän.CSS.44
2437	förrän.S	26.008873	9	förrän.SPS.29
2451	bara.CS	25.793389	7	bara.CSS.34
2530	innan.S	24.473397	8	innan.SPS.27
2598	för.CS	23.623399	7	för.CSS.34
2606	vissa.PI	23.541650	8	vissa.PI@0P0@S.26 wissa.PI@0P0@S.1
2615	förbi.S	23.451727	6	förbi.SPS.38
2698	oavsett.S	22.339481	7	oavsett.SPS.27
2768	få.PI	21.360868	8	få.PI@0P0@S.29
2779	varannan.DI	21.184624	8	vartannat.DI@NS@S.7
	varannan.DI@US@S.18			
2809	ens.PS	20.740231	7	ens.PS@000@S.27
2810	densamma.PF	20.691895	6	densamme.PF@MS0@S.3
	densamma.PF@US0@S.30			
2819	igenom.S	20.621269	7	igenom.SPS.27
2846	varenda.DI	20.301893	5	varenda.DI@0S@S.26
	hvarenda.DI@0S@S.2			vartenda.DI@NS@S.2
2875	bägge.DF	20.039643	7	bägge.DF@0P@S.30
2898	varandras.PS	19.825858	7	varandras.PS@000@S.25
2944	liksom.CS	19.264917	8	liksom.CSS.22

FB4– Function Words

3014	bortom.S	18.680443	7	bortom.SPS.24
3138	dennas.PS	17.430748	8	dennes.PS@000@S.17
	dennas.PS@000@S.3			dennes(as).PS@000@S.1
3144	rörande.S	17.382643	6	rörande.SPS.29
3205	jämte.S	16.935201	8	jämte.SPS.20

3234	sen.S	16.653262	6	sen.SPS.27
3255	utom.CC	16.514412	6	utom.CCS.24
3347	innanför.S	15.593448	6	innanför.SPS.25
3419	alltihop.PI	15.106955	6	alltihop.PI@NS0@S.22
3483	somlig.PI	14.688325	7	somliga.PI@0P0@S.19
	somligt.PI@NS0@S.1			
3532	uppför.S	14.334917	4	uppför.SPS.32
3598	ifall.CS	13.844536	4	ifall.CSS.29
3637	intet.PI	13.613484	8	intet.PI@NS0@S.18
3640	samtliga.PI	13.590827	7	samtliga.PI@0P0@S.16
3645	utmed.S	13.547583	5	utmed.SPS.19
3698	nedanför.S	13.178174	6	nedanför.SPS.20
3820	ovanpå.S	12.417175	5	ovanpå.SPS.20
3897	allas.PS	12.061277	6	allas.PS@000@S.16

FB5– Function Words

4095	utefter.S	11.037995	6	utefter.SPS.16
4097	allteftersom.CS	11.035816	6	allteftersom.CSS.15
4134	angående.S	10.852088	4	angående.SPS.24
4187	såvitt.CS	10.581776	7	såvitt.CSS.13
4245	vardera.PF	10.276956	5	vardera.PF@0P0@S.17
4373	varsin.PS	9.689170	6	varsitt.PS@NS0@S.1
	varsin.PS@US0@S.15			
4447	alltsammans.PI	9.390133	4	alltsammans.PI@NS0@S.19
4471	bägge.PF	9.296207	5	bägge.PF@0PS@S.14
4526	sen.CS	9.115134	4	sen.CSS.19
4603	ömsom.CC	8.735467	4	ömsom.CCS.15
4678	uppåt.S	8.485058	5	uppåt.SPS.14
4682	alltsedan.S	8.477270	7	alltsedan.SPS.12
4698	invid.S	8.401518	4	invid.SPS.17
4850	inifrån.S	7.873816	5	inifrån.SPS.13

FB6– Function Words

5049	inuti.S	7.208002	4	inuti.SPS.13
5087	alltifrån.S	7.092017	6	alltifrån.SPS.10
5120	allihop.PI	7.017216	5	allihop.PI@0P0@S.9
	allihopa.PI@0P0@S.2			
5738	intet.DI	5.586077	6	intet.DI@NS@S.10
5741	envar.PI	5.577443	4	envar.PI@US0@S.9
5772	inpå.S	5.534193	5	inpå.SPS.10
5935	vardera.DF	5.150550	4	vardera.DF@0S@S.9

FB7– Function Words

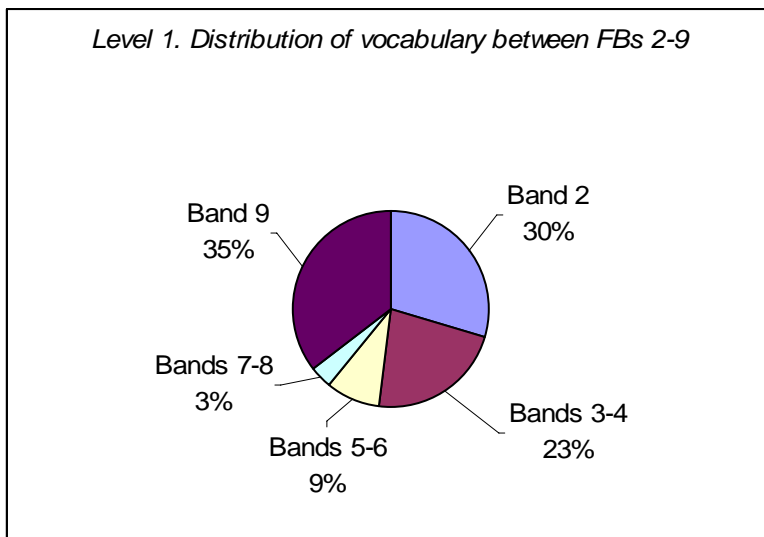
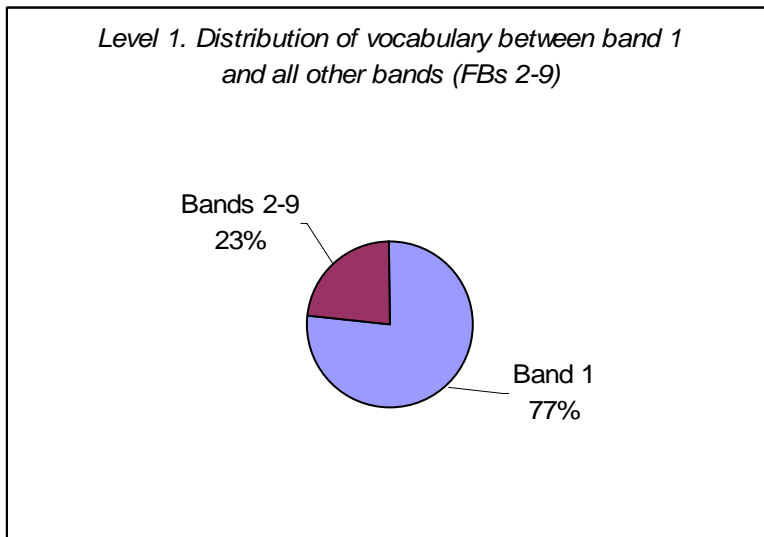
6125	vilkas.PE	4.733101	4	vilkas.PE@000@S.9
6263	undan.S	4.474224	4	undan.SPS.11
6378	någons.PS	4.272436	5	någons.PS@000@S.7 nåns.PS@US0@S.1
6926	varannan.PI	3.350121	4	vartannat.PI@NS0@S.4
	varannan.PI@US0@S.1			

FB8– Function Words

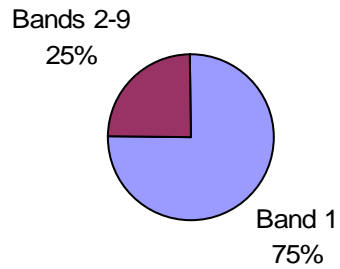
7027	uppemot.S	3.177323	4	uppemot.SPS.6
7281	desamma.PF	2.737300	4	desamma.PF@0P0@S.7
7367	såvida.CS	2.581154	4	såvida.CSS.5
7532	alltmedan.CS	2.253247	4	alltmedan.CSS.5

Appendix 3. Diagrams of FB distribution per each LIX level, average values

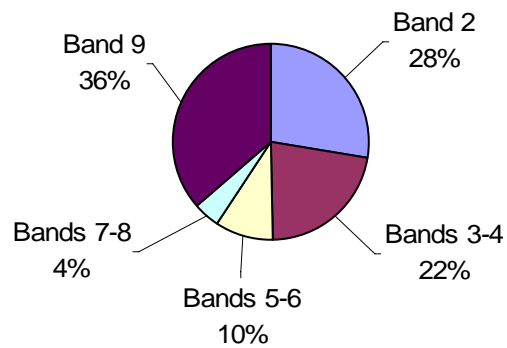
The first diagram for each LIX level shows the (average) distribution of vocabulary from FB1 as contrasted to any other vocabulary (FBs 2-9). The second diagram shows the distribution between vocabulary from bands 2-9.



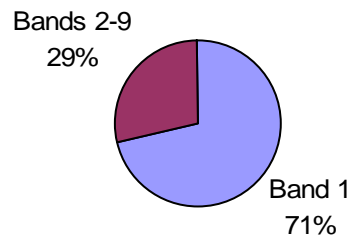
Level 2. Distribution of vocabulary between band 1 and all other bands (FBs 2-9)



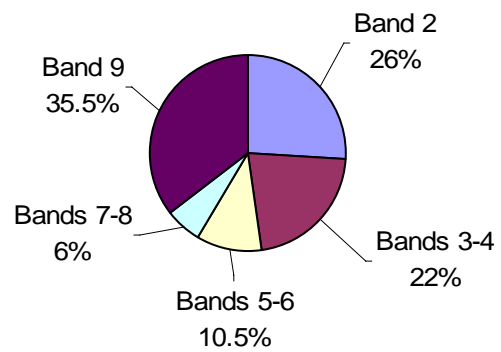
Level 2. Distribution of vocabulary between FBs 2-9



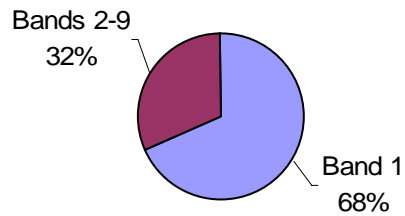
Level 3. Distribution of vocabulary between band 1 and all other bands (FBs 2-9)



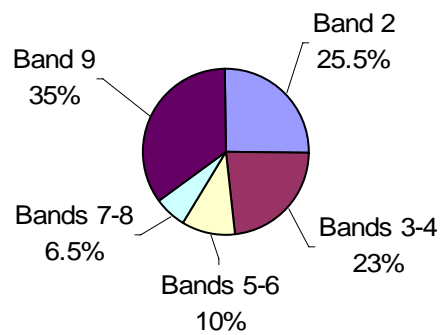
Level 3. Distribution of vocabulary between FBs 2-9



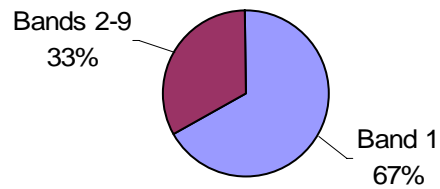
Level 4. Distribution of vocabulary between band 1 and all other bands (FBs 2-9)



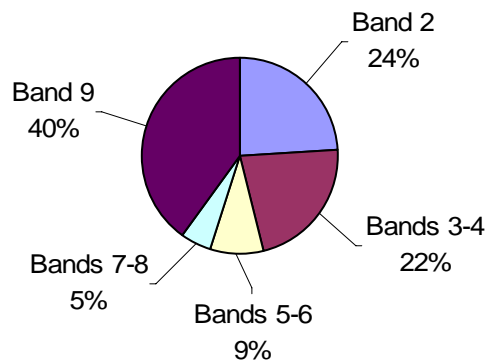
Level 4. Distribution of vocabulary between FBs 2-9



Level 5. Distribution of vocabulary between band 1 and all other bands (FB 2-9)



Level 5. Distribution of vocabulary between FBs 2-9



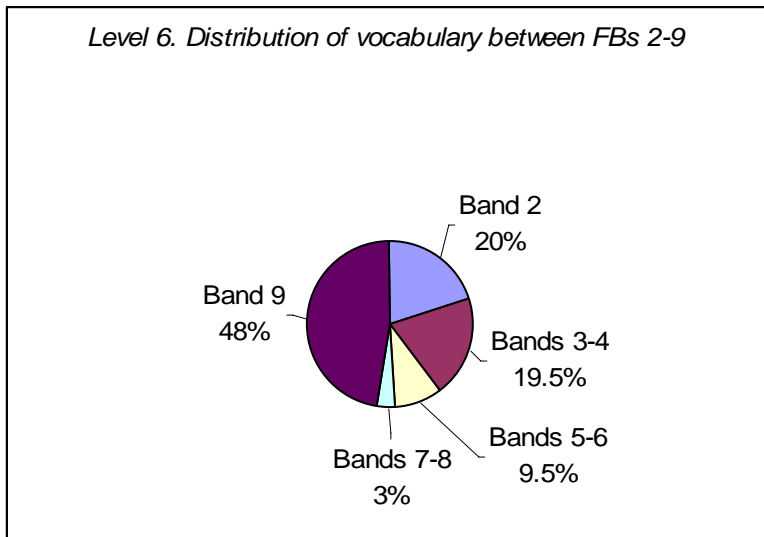
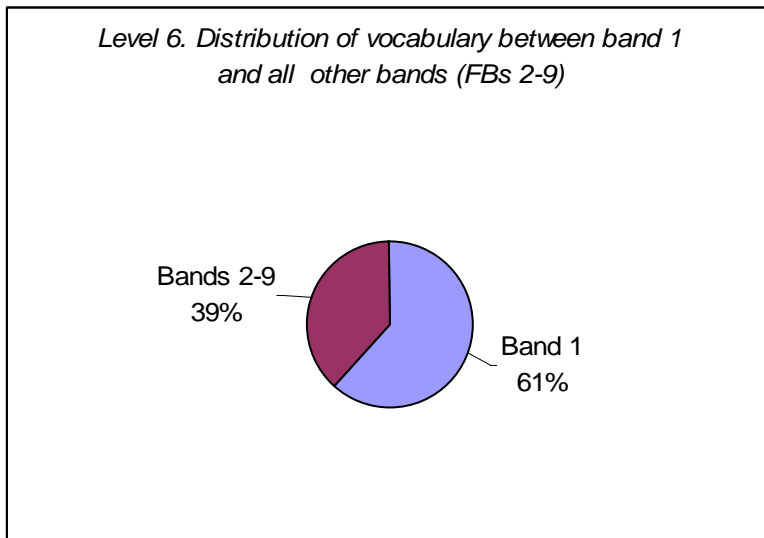


Table 16. Standard deviation of FB, LD and LV values per each level.

Standard deviation	Band 1	Band 2	Bands 3-4	Bands 5-6	Bands 7-8	Band 9	LD	LV
Level 1	3.5	4.9	3.2	1.9	1.2	6.1	2.8	3.5
Level 2	3.9	4.1	3.3	2.1	1.6	5.8	3.4	2.6
Level 3	5.4	4.6	3.6	2.2	2.1	6.7	3.1	4.9
Level 4	4.7	4.4	3.9	2.8	2.6	7.3	3.0	6.0
Level 5	5.2	5.6	4.6	2.5	2.2	9.1	3.7	5.3
Level 6	4.0	3.2	3.4	3.6	1.2	9.6	4.4	3.2

Appendix 4. Texts used for readability grading by human readers

The texts below have the same names as in SUC database. They start with letters that identify the genre followed by the running number of the text within the genre. The symbol @ stands for new paragraphs, though not consistently from text to text. Each text is preceded by statistics on its vocabulary frequency profile, LD, LV values as well as LIX, LexLIX and LFP scores.

<<<ha23>>>

LIX values	Band1, %	FBand2, %	FBands 3-8, %	FB9, %	LFP score	Lex. Density	Lex. Variation	LexLix
38	50.1	10.0	12.2	27.5	77.4	60	31	57.7

@@@ Lag om avgifter på vissa jordbruksprodukter m. m.;

@utfärdad den 14 juni 1990.

@ Enligt riksdagens beslut föreskrivs följande.

@@Avgifter vid införsel

@ 1 #pp För att skydda priserna på jordbruksprodukter får regeringen föreskriva att en avgift skall tas ut på varor som förs in i landet. Sådan avgift får tas ut på varor som anges i bilagan till denna lag.

@ För samma ändamål får regeringen föreskriva att sådana varor som anges i bilagan till denna lag inte får föras in i eller ut ur landet utan särskilt tillstånd.

@ 2 #pp En avgift enligt 1 #pp skall debiteras och uppbäras av tullverket i den ordning som gäller för tull. Även i övrigt gäller vad som är föreskrivet om tull.

@ Regeringen får meddela föreskrifter om undantag från första stycket.

@@Avgifter för att utjämna industrins råvarukostnader

@ 3 #pp För att bekosta en utjämning av industrins råvarukostnader får regeringen föreskriva att en avgift skall betalas för varor som tillverkas av råvaror som anges i bilagan till denna lag. Sådan avgift får tas ut på varor som tillverkas inom landet för försäljning eller som förs in i landet och förtullas.

@ 4 #pp I fråga om avgifter enligt 3 #pp för varor som importeras tillämpas 2 #pp.

@ Avgifter enligt 3 #pp för varor som tillverkas inom landet skall debiteras och uppbäras av riksskatteverket i den utsträckning regeringen föreskriver.

@@Tillverkningsavgifter på fettvaror

@ 5 #pp För att reglera priset på fettvaror får regeringen föreskriva att tillverkare skall betala avgift vid tillverkning av vegetabiliskt och animaliskt fett.

@ 6 #pp Regeringen får föreskriva att tillverkare skall betala avgift för foder som framställs i samband med att olja utvinns ur vegetabiliska råvaror.

@@Utbetalning av prisstöd

@ 7 #pp Slakteri, fristående sanitetsslaktavdelning och mejeri skall betala ut prisstöd av medel som statens jordbruksnämnd ställer till företagets förfogande. För prisstöd som betalas ut skall redovisning lämnas till jordbruksnämnden.

<<<kk59>>>

LIX values	Band1, %	FBand2, %	FBands 3-8, %	FB9, %	LFP score	Lex. Density	Lex. Variation	LexLix
------------	----------	-----------	---------------	--------	-----------	--------------	----------------	--------

21	77.6	7.9	7.4	7.1	29.5	50	26	25.2
----	------	-----	-----	-----	------	----	----	------

@Jag ville göra något extra för Matti. Jag tog med honom till en hundfrissa. Men han var bara rädd för saxarna och han darrade när han tvingades stå högt uppe på ett bord. Det blev så fel. Jag ville visa, både för Matti och mig själv, att jag tog mig tid med honom, även nu.@Matti blev fin till slut. Speciellt runt öronen. Han hade fått page-frisyr och såg ut som en liten flicka.

@Det mesta fanns i mig. Jag kände mig som en skurk när jag lyfte över Sigge till mig på nätterna och han fick sova i min säng. Matti tittade, mer var det inte. Han protesterade inte. Men för mig räckte det att han tittade. Jag kände hur jag svek honom. Jag förstod att han visste att jag hade någon annan att bry mig om nu, någon som jag värderade högre än honom. Någon med mer omedelbara behov än vad Matti hade.

@Vintern och våren gick. Jag tog promenader med Matti som vanligt. Sigge var med. Han satt i en sele på min mage och vaggades till sömns av mina steg. Matti fick rostat bröd morgon och kväll. Jag kokade ris och blandade i hans mat. Jag ville till varje pris övertyga Matti och mig själv om att allting var som vanligt, nästan. Men i mitt huvud kretsade tankarna kring Sigge.

@När Sigge sov middag försökte jag göra inomhusövningar med Matti. Sök, lydnamdsmoment. Men jag var trött av att amma och vara vaken på nätterna. Det tog emot ännu mer att träna Matti när jag visste att han bara blev mer uppjagad efteråt.

@Ibland orkade jag inte. Då blev jag istället sittande vid köksbordet och räknade de många timmar som Matti skulle vara inaktiv det dygnet. I hallen eller ute på bron låg Matti och väntade. Mitt dåliga samvete växte. Det tog sig olika uttryck. Vissa gånger i överdriven aktivitet på veckosluten. Då var Janne med Sigge och jag kunde ta mig an Matti långa stunder. Andra gånger i resignation. Jag gick och lade mig och sov bort dagen. När kvällen kom var den dagen ändå förlorad. Jag hade försummat Matti, men det var för sent att göra något åt det och på ett märkligt sätt var den tanken en befrielse: det är ändå för sent.

@Vad är jag för en människa som fick mitt hundägande att bli till tvångsmässiga ritualer? Berodde det bara på att Matti var den hund som han var? Blev han till den hund som han var på grund av mig?

@På sommaren var Matti med extrafamiljen i deras sommarstuga några dagar. Janne och jag tog ut en filt i trädgården. Vi ställde ut en balja i gräset och badade Sigge. Det var sådant som inte gick att göra när Matti var hemma. Han skulle ha slitit i filten. Han skulle ha ställt sig bredvid Sigge och skällt när han såg att Sigge fick bada. Han skulle ha kommit med bollar och pinnar och krävt att vi skulle kasta dem. Han skulle ha trängt sig in på oss, bokstavligen. Han gjorde så om man satt på golvet inne eller i gräset ute. Kom och trängde sidan av kroppen mot en. Satte sig i ens knä. Om man sa till honom att lägga sig bredvid istället gjorde han kanske det. I fem sekunder. Sedan reste han sig och flämtade och stirrade eller sprang iväg och hämtade en leksak. Om vi stängde in honom skällde han oavbrutet.

@Vi hämtade hem Matti från extrafamiljen. De berättade att han hade jagat en bil som hade kört upp på gården och hoppat mot den så att det hade blivit repor i lacken.De var förvånade. De hade aldrig sett Matti uppföra sig så.

<<<fa02>>>

LIX value	Band1, %	FBand2, %	FBands 3-8, %	FB9, %	LFP score	Lex. Density	Lex. Variation	LexLix
35	82.2	7.0	7.4	3.4	21.2	55	29	28.1

@ @ @Samtalet och identiteten

@En av mina studenter, Lena, har varit hemma några veckor för att skriva färdig sin uppsats. Hon kommer in på mitt rum för att lämna in den och jag frågar hur hon haft det. Frågan möts först av tystnad. Hon har tagit den på allvar och ser bekymrad ut. Hennes svar kommer trevande och de dialektala dragen - som jag förut knappt lagt märke till - är tydliga. "Dom tyckte hemma att jag börjat tala konstigt, jag kände ibland, ja, jag kände mej främmande liksom, som jag inte hörde dit. Jag hörde själv att jag använde ord som, ja, nya ord. Tillgjort lät det. Jag visste ibland inte vem jag var. Mitt språk hade förändrats. Men jag är ju densamma. Eller?"

@De flesta moderna utbildningar ger oss ett delvis nytt språk. Vi får nya ord för både gamla insikter och nya begrepp. Vi anammar nya sätt att uttrycka oss som är typiska för den grupp vi tillhör eller vill tillhöra. Förändrar vi vår identitet när vi förändrar vårt språk?

@Det nya språk och de nya ord som Lena tagit med sig hem från universitetsvärlden och plötsligt hört med nya öron var inte bara redskap för tanken. Lena, som är en tänkande och djupt engagerad student hade kanske hemma börjat tala om "insocialisering", "attityder" och "struktur". De här orden, liksom alla andra ord, bar med sig de sammanhang de använts i. För föräldrar och släkt vittnade de om en ny miljö med främmande värderingar och de knappt hörbara uttalsförändringarna gav besked om nya intressegemenskaper. Lena hörde plötsligt själv hur hennes språk hade ändrats och kände sig förvirrad. Svek hon sitt ursprung, förnekade hon sin samhörighet?

@Vi vet att individen växer och formas genom att upptas i dialog med andra. Genom att samtala mejslas vårt jag fram. Vår identitet är med andra ord starkt förknippad med vårt språk och våra möten med andra. Språket avspeglar vår tillhörighet - den geografiska, yrkesmässiga och sociala. Det avspeglar vårt temperament - livligt eller eftertänksamt. Det bär hemligt våra innersta känslor och tankar, men det kan också förråda oss och avslöja oss. Det kan röja vår identitet, det kan visa vilka vi är.

@Ibland kan det hända att man befinner sig i en situation tillsammans med människor där man inte känner sig hemma. Det är svårt att samtala. Ens "jag" passar inte in, ens person och språk verkar inte accepteras. Man känner sin identitet hotad. Vad händer då? Antingen finner man sig i detta och förblir utanför eller också anpassar man sig till situationen och spelar med så gott det går. Det kan till synes gå bra, men man kan ändå känna sig konstig. "Det här är inte jag" tycker man. Det kan också hända att man nästan skäms - man har undertryckt sin identitet, eller åtminstone någon del av den.

@Vad är vår identitet? Det här är en svår fråga och här får det räcka att säga att svaret måste sökas i två andra frågor: "Vem är jag?" och "Vem är jag för andra?". Identiteten är dels den aspekt av oss som vi uppfattar som typisk för just oss själva, dels den aspekt som vi tror andra uppfattar som typisk. Det finns socialpsykologer som påstår att identiteten är ett svar på alla de situationer vi ställs inför och att vi därför uppvisar flera jag. Vi reagerar olika inför olika situationer och människor och alla dessa reaktioner utgör delar av vår individuella personlighet, även om de skulle vara motsägelsefulla.

<<<kk13>>>

LIX values	Band1, %	FBand2, %	FBands 3-8, %	FB9, %	LFP score	Lex. Density	Lex. Variation	LexLix
23	80.2	6.8	6.4	6.7	26.5	55	32	24.7

@@Ariel

@När Anna Davenport låg på förlossningsbordet och pressade ut sitt första barn kände hon något lent och främmande röra lårens insidor. Mitt inne i smärtan fanns någonting som gjorde ofantligt gott. Sol kom genom de stora fönsterrutorna och hon drog in en doft av sädesfält och lindblom, som inte fanns utanför de här fönstren, utan långt borta, vid det hus hon själv var född i.

@ Och Anna böjde på nacken; en gammal barnmorska höll barnet i sina armar och då - just när kvinnan svängde barnet mot henne såg Anna att det var en flicka, och en flicka med vingor hon fött till världen.

@ Vingarna var så tunna att bara solens strålar lyckades reflektera dem, en tunn skimrande blå hinna och Anna skrek till i rädsla över att barnmorskan skulle skada dem.

@ - Akta vingarna!

@ Den gamla kvinnan var bred över höfterna och stel i kroppen och hon måste böja sitt ansikte tätt intill barnets rygg.

@ - Ja, sannerligen har hon inte det! Då ska du se att hon kommer att flyga.

@ Och hon lade barnet till Annas bröst.

@ Sen satte hon sig på sängkanten:

@ - Säg mig, flicka lilla, har du en bra man?

@ - Vad menar du? Anna försökte le. Nog har jag väl det.

@ - Då så, sa gumman.

@ - Men du menade något särskilt?

@ Gumman ryckte på axlarna:

@ - Nej, nej, nu ska vi sova, sa hon och i samma stund föll Anna i en så djup sömn att hon drog hela dagen med sig.

@ Filip hade kommit mitt under maskrosskörden, den dag i maj när Anna och hennes systrar samlade blommor till årets vinberedning. Detta var en av årets bästa dagar. Den betydde att sommaren var här och att vattnet i pumpen porlade på ett särskilt sätt när de då och då måste dit för att tvätta bort den sträva maskrosmjölken. Den betydde att få gå fram till pappa, som satt på den bruna bänken under äppelträdet, tömma korgen och få en klapp på kinden.

@ Tillsammans med två andra pojkar stod han plötsligt inne i trädgården, i träningsoverall, med ryggsäckar och tält och undrade om det fanns plats en natt eller två på ägorerna. De var amerikaner, hade skägg och glasögon, studerade vid universitetet, nu ville de lära känna landet. Det var den längste av dem som talade, och när han talade, tänkte Anna, hörde man först grammatiken, sedan orden.

@ Framemot kvällen rodde de sex ungdomarna ut på sjön. Det var den länge som rodde och Anna hade hamnat mitt emot honom. Det kanske blir så, tänkte hon långt senare. Att den som sätter sig mitt emot någon vid ett visst tillfälle, den blir det.

@ - Akta dig för grynnan där, pekade hon.

@ - Grynna? Jag känner inte betydelsen av det ordet.

@ Som om han höll sig i grammatikens ledstång. Men han vände sig om:

@ - Grynna. Någoting lågt och förrädiskt, jag förstår. Grynna.

@ Den natten sov inte Anna. Hon satt i sitt fönster och tittade i kikaren ner mot tältet vid sjön. Hon kunde bara se den länge, allvarlige. Med korslagda ben satt han utanför tältöppningen och läste i en bok, medan handen då och då jagade iväg en mygga. Och när hon tidigt om morgonen gick ut för att kissa, såg hon honom fortfarande sitta på samma sätt, men huvudet hade fallit framåt, han sov, boken låg i gräsets dagg. Barfota i det blå nattlinnet smög hon ner genom hagen. De små lövgrodorna hoppade åt sidan, den mjuka leran pressade sig upp mellan hennes tår. Tyst plockade hon upp boken och tryckte den, medan hon sprang upp mot huset, tätt till sitt bröst. Boken hette Naturen och Människan av Ralph Waldo Emerson. Anna läste. Hon läste fortfarande när de första fåglarna tjattrade i plommonträdet, ovan vid språket, vid bilderna, men det var inte för sig själv hon läste: hon ville genom Emersons ord se vem han var, han som nyss och så ivrigt läst. Hon behövde förstå i vilket sorts ljus han levde. Hon hade visserligen svårt att tro på det ljuset, när hon såg hans allvarligt svarta blick, men man kan inte se allt.

<<<k19>>>

LIX values	Band1, %	FBand2, %	FBands 3-8, %	FB9, %	LFP score	Lex. Density	Lex. Variation	LexLix
37	70.7	6.5	12.0	10.7	40.0	54	39	38.5

@ - Jag är ledsen, sa hon lågt. Jag kan inte tala om det. En annan gång kanske. Jag heter Rut Gren förresten och jag bor alltså på Själagårdsgatan. Och jag finns i telefonkatalogen. Hon log.

@ - Jag heter Johan Homan, sa jag. Och jag är antikhandlare. Granne faktiskt, för jag har min affär på Köpmangatan.

@ - Är det ni? sa hon förtjust. Jag som så ofta går förbi er affär. Ni har alltid så smakfull skyltning. Men jag vågar aldrig gå in. Min pension tillåter inte några utsvävningar på antikmarknaden.

@ - Välkommen. Det är inget köptvång. Och allting kostar inte skjortan.

@ - Nästa gång jag går förbi så tittar jag in. Fast då får jag väl lämna Jussi hemma så att han inte ger sig på lilla Cl'o.

@ - Eller tvärtom.

@Rut Gren skrattade och skulle just säga någonting. Men vad det var fick jag inte veta för sköterskan öppnade dörren och bad henne stiga in.

@Jag satt kvar i min fåtölj och såg efter henne och hennes lilla trasselsudd till hund. Vilket människoöde! Hur var det möjligt att leva vidare efter hennes upplevelser? Vilken oerhörd styrka och livskraft hon måste ha för att kunna överleva lägren. Och tiden efteråt. Om jag hade blivit oerhört illa berörd av vad jag sett på TV och i tidningarna av skrånande gatumobb med antisemitiska skymford, hitlerhälsningar och nazistflaggor så måste det vara tusen gånger värre för henne. Hon hade sett det hända, varit där när skyltfönster till judiska affärer slogs in. När synagogor brändes och när oskyldiga människor slogs ner på gatorna inför skrattande poliser. Hon hade sett dödsänglarna i koncentrationslägren, sett hur människor förintats. Och hon hade bevittnat hur cancern växte för att till slut förtära ett folk, ett land, en stat där med grym ironi den väg som ledde till Belsens dödsläger hette Beethovenstrasse.

@Cl'o jamade förebrående i sin bur. Jag öppnade locket, tog upp henne i mitt knä där hon spinnande rullade ihop sig. Faran var över. Den otäcka hunden, vilddjuret från Skottland, den sentida pocketversionen av Baskervilles hund hade försvunnit.

@Kunde hon haft rätt, den lilla späda damen? Kunde hon verkligen ha sett en av bödlarna från Auschwitz i Gamla stan femtio år efter kriget? Och det slog mig plötsligt hur väl hennes beskrivning stämde in på doktor Wagner. Kunde doktor Wagner vara Gamen? SS-dolken i hans bröst, var den en hämndaktion från dem som inte kunnat få rättvisa genom lagens långsamma och ineffektiva maskineri där kvarnarna malde tomt för krigsförbrytare? Och jag tänkte på Wagners egendomliga, svarta ögon. Auran av ensamhet, kyla och mörker som omgav honom. Den instinktiva olust och nästan avsmak jag känt inför hans mörka gestalt när han stått utanför min dörr. Men jag inbillade mig naturligtvis. Hade Wagner varit en så prominent förgrundsfigur i nazisternas dödsläger så måste han väl spårats för länge sedan? Och även om han lyckats hålla sig undan i alla dessa år så vad gjorde han i så fall i Gamla stan? Vad hade han för anledning att hyra en våning i mitt hus? Nej, sanningen var väl att den stackars kvinnans fruktansvärda upplevelser satt så djupa spår att hon inte kunde göra sig fri, att hon fortfarande reds av maran om nätterna. Så drömmer hon om det där monstret och det första hon ser när hon kommer ut på morgonen är en svartklädd man som påminner om Gamen. Ett hjärnsnöke ur en stackars lidande kvinnas skräckdrömmar var väl den troliga förklaringen.

<<<gb17a>>>

LIX values	Band1, %	FBand2, %	FBands 3-8, %	FB9, %	LFP score	Lex. Density	Lex. Variation	LexLix
41	68.7	8.1	14.6	8.5	39.8	57	41	40.4

@ @ @ Vaclav Havel

@ @ @ EN FILOSOF PÅ TRONEN

@ HANS ISAKSSON

@ I USAs krig mot Irak ser han en "sund självbevaringsdrift". Den kapitalistiska äganderätten finner han "naturlig". När han ljuger för folket tycker han det är äckligt - men fortsätter likväl.

@ Tjeckoslovakiens president är ett helgon i tiden.

@ Vaclav Havel tillhör svenska kultursidors absoluta favoriter. Han konsulteras ofta i aktuella världsfrågor. Likt alla stora filosofer kan Havel allt om intet. Men kan yttra sig om allt. I DN 12/1 är det Adam Michnik som sitter på mästarens knä och intervjuar. Dagens audiens gäller i huvudsak trenne ämnen: Uppgörelsen med den gamla regimens anhängare i öststaterna, ideologiernas framtid och hur Havel själv mår.

@ Havel får först motivera den nya lagstiftning som i praktiken innebär yrkesförbud för alla f d partifunktionärer i forna CSSR från kommunnivå och uppåt. Personligen säger sig den alltid lika blide som tankspridde Havel ha slarvat bort lappen med namnen på alla författarkollegor som en gång angivit honom.

Det låter som en förtjusande förevändning för utövande av storsinnet. Men som statsman kan han tyvärr inte vara lika distraherad, suckar Havel - folket kräver sin hämnd och måste få det. Och vem kan väl säga däremot?

@Dock - varför nöja sig med att avskeda folk, om de verkligen förbrutit sig mot någon lag, eller mot mänskligheten? Varför inte rannsaka, döma och bestraffa de skyldiga som individer?

@Havel fruktar ett återfall i totalitarism. Det finns en grogrund för "dem som drömmer om en stark man som ställer allting till rätta. Och det är nog mindre viktigt för dem, om denne man viftar med vänsterns eller högerens fana".

@Själv är Havel inte så orolig för högerfanan.

@Idag är bara höger på modet, helt enkelt, begripligt av flera skäl, antar jag. Vad annat hade man kunnat vänta sig efter kommunismens fall? Det är helt enkelt en normal motreaktion".

@Nej, menar Havel, folk borde inse att ideologierna är döda och att vi stigit in i idernas epok. Hotet kommer nu från dem som vägrar att inse det nya läget, dvs från fundamentalismerna av olika slag - religiösa, nationalistiska, ideologiska. Man borde inränga kommunismen här också, postumt. Så tycker Andr` Glucksmann. Och Glucksmann tillhör också den nyuropeiska filosofparnassen. Havel känner sig befryndad, men nyanserar, förstås:

@Jag delar Glucksmanns oro, men ser också en motkraft som jag hoppas kommer att segra. Jag tänker på världens självbevaringsdrift. Lägga märke till att när Saddam Hussein överföll Kuwait reagerade för första gången hela världssamfundet, inklusive arabstaterna och FN. Det var något nytt som jag tolkar som utslag av sund självbevaringsdrift. Till synes gällde det en liten stat med några borrhögar. Men på spel stod risken för utbredning av det fundamentalistiska vansinnet, hot mot andra stater och folkmord mot kurderna till att börja med. Det såg ut som om mänskligheten började bli medveten om allvaret i detta hot, i annat fall hade Bush och Baker inte lyckats, inte ens om de varit hundra gånger smartare."

@Man kan rikta invändningar mot sakframställningen, och även ifrågasätta om Havels framställning av motiven för USA-alliansens intervention mot Irak är särskilt uttömmande, eller ens rättvisande, men vem kan ifrågasätta dess "självbevaringsdrift"?

@Det finns alltså sunda motkrafter i det kollektiva medvetandet, och det är Havel förunnat att skönja, och för världen förkunna dem. När Libyen genom illa dolda hot om nya terrorbombningar skall tvingas på knä, när vi nås av meddelanden om att USA:s flotta står i begrepp att borda Nordkoreanska handelsfartyg, som kunde vara på väg till Iran och ha Scudmissiler ombord hör vi samtidigt från tongivande kretsar i USA uttryck för tillfredsställelse över att hyckleriet med de små, ofta oansvariga, nationernas sk suveränitet i och med interventionen i Jugoslavien äntligen brutits. Då vet vi, tack vare Havel, att det inte rör sig om USA:s kanonbåtsdiplomati eller stormaktsarrogans, utan om den civiliserade världens naturliga självbevaringsdrift. Havels egen nation är visserligen liten, och kommer att bli än mindre när han är färdig med sin utförsäljning till Stortyskland, och styckat av Slovakien, men är ändå hela tiden oerhört ansvarig.

<<<bb01a>>>

LIX values	Band1, %	FBand2, %	FBands 3-8, %	FB9, %	LFP score	Lex. Density	Lex. Variation	LexLix
44	69.7	8.7	13.9	7.7	38.0	58	38	41

"En show på tittarnas bekostnad"

Aktuellt nyhetsrapportering döms ut i ny undersökning av elever vid Journalisthögskolan.

Aktuellt bevakning av vinterns lärarkonflikt var skrämmande skev. Makthavarna lyckades utnyttja TV-medietets förkärlek för kända ansikten och aggressiva utspel. SACO favoriserades både innehålls- och tidsmässigt i den subjektiva rapporteringen. Det visar en undersökning av Ulrika Krebs och Ia Röhl, studenter vid Journalisthögskolan. Nyhetsredaktionen har inte förstått sin viktiga funktion i samhället. Bevakningen har blivit en show, skriver de.

En öppen och direkt kontakt mellan makthavare och journalister är något som vi förknippar med demokratier. Det skriver Ekots inrikeschef Thomas Hempel (DN Debatt 12/3). Ja, det låter som en självklarhet. Men när han kopplar öppenheten till att "stå i korridorer" eller att "vara på plats och förmedla röster och stämningar", så riskerar det att bli en sådan tom uppvisning som Stig Hadenius talar om på DN Debatt 28/2. En uppvisning utan innehåll, utan svar på våra frågor. Det finns en motsättning mellan patrullerandet i korridorer och annan form av journalistik, först och främst av den enkla anledningen att både resurser och sändningstid är begränsade.

Vad är det då som försvinner i valet mellan olika arbetsmetoder och värderingar? Vi har granskat samtliga 95 inslag i Aktuellt's bevakning av lärarkonflikten, mellan den 31 oktober och den 15 december 1989, och ska inom kort presentera våra resultat vid Journalisthögskolani Stockholm. När man ser rapporteringen i sin helhet, så är det en skrämmande skev bild av verkligheten som Aktuellt valt att visa oss.

Konflikten gällde lärarnas oenighet i avtalsrörelsen; TCO, som representerar låg- och mellanstadielärare och i viss mån även högstadielärare, svarade ja till arbetsgivarens bud eftersom det innebar stora löneökningar för medlemmarna. SACO, som representerar gymnasielärarna och en stor del av högstadielärarna, sa däremot nej till budet, därför att det innebar en ökad närvaroplikt och lika slutlön för skilda lärargrupper.

När Stig Hadenius (DN Debatt 28/2) klagar på att nyhetsförmedlingen domineras av "partipolitiskt käbbeloch gräl mellan arbetsmarknadens parter", tröstar han sig ändå med att den som regel är objektiv - "de behandlar olika parter på ungefär samma sätt". Vår studie visar på motsatsen. SACO-lärarna som tog till stridsåtgärder kom att dominera hela rapporteringen.

En strejk har naturligtvis högt nyhetsvärde, eftersom den drabbar andra grupper, men det avtal som skulle slutas med statens arbetsgivarverk gällde samtliga lärare, och därför borde inte SACO-lärarna så självklart ha fått spela huvudrollen. Tittar vi på de olika fackliga företrädnas och lärarnas medverkan blir bilden av partiskhet tydlig. 38 lärare totalt uttalade sig för eller emot arbetsgivarens bud, av dessa var 29 kritiska och stödde SACO:s linje (i tid 7 minuter och 38 sekunder), medan endast 9 lärare uttryckte sitt stöd för jasiidan, det vill säga TCO:s linje (i tid 2 minuter och 38 sekunder). Ove Engman, SACO-lärarnas facklige företrädare, fick tala i sammanlagt 19 minuter och 24 sekunder. De två fackliga företrädarna för TCO, Solveig Paulsson och Christer Romilsson, fick sammanlagt tala i 5 minuter och 4 sekunder.

<<<bb05a>>>

LIX values	Band1, %	FBand2, %	FBands 3-8, %	FB9, %	LFP score	Lex. Density	Lex. Variation	LexLix
45	70.8	6.8	11.6	10.7	39.9	59	41	42.5

Ge ungdomarna politisk makt

Moderat ungdomsordförande vill ha bort pensionärerna som blockerar de politiska uppdragen i nämnderna

UNGDOMARNA ÄR UTESTÅNGDA från politiska uppdrag. De äldre politikerna släpper inte ifrån sig makten. Det hävdar Fredrik Reinfeldt, ordförande för ungdomsmoderaterna i Stockholm, i ett inlägg om det dåliga intresset för kommunalpolitik. Peter Hellsten, Tyresö, påstod i en debattartikel den 28 mars att den politiska aktiviteten nu är så låg att "mycket små grupper av beslutsamma människor närmast kuppattat kan ta över beslutanderätten i många fullmäktigeförsamlingar".

Peter Hellsten skriver på Stockholmsdebatt 1990-03-28 att det är massmediernas fel att ungdomar flyr från partiernas möten. I mångt och mycket tycker jag att Peter Hellsten tecknar en riktig bild av dagens politiska verklighet. Han har dock glömt en viktig sak. Det är inte alla ungdomar som har "flytt in i cocacola-kulturen", de politiska ungdomsförbunden har trots att de tappar medlemmar fortfarande stor attraktionskraft. Problemet är bara att för de flesta slutar engagemanget efter gymnasieåldern. Efter fyllda 20 avfolkas ungdomsförbunden snabbt på intresserade som kan tänka sig att ta kommunalpolitiska uppdrag. Peter Hellsten förklarar detta med att massmedierna skrämmer bort dem, men de möts också av ett politiskt system och ett nomineringsförfarande inom partierna som stoppar dem.

Politiken i dag tillhör dem som redan har förtroendeuppdrag och de är inte intresserade av att dela med sig av uppdragen, framför allt inte till ungdomar. De politiska partierna saknar en strategi för hur man skall hjälpa unga människor in i politiken och därmed föryngra de beslutande församlingarna.

Tydligast syns detta i Stockholms stad.

I Stockholm uppmärksammas politikerna av massmedierna betydligt mer än i mindre kommuner. Detta gör det intressant att vara politiker. Det politiska arbetet är dessutom helt anpassat för att gynna pensionärer och offentliganställda med stora möjligheter att ta ledigt.

Nämnderna sammanträder på kontorstid och konferenser och dylikt läggs mitt i veckorna. Fritidspolitikerna i Stockholm, som borde finna det svårt att hinna med en enda av stadens tyngre nämnder, sitter ofta i två eller till och med tre tyngre nämnder. Varje nämnduppdrag leder oftast till en mängd mindre följduppdrag.

Stockholm betalar fasta årsarvoden för nämnduppdragen vilket gynnar dem som tar på sig många uppdrag. Det finns förtroendevalda som tjänar över 5000 i månaden på fasta arvoden, utan att ha närvarat vid ett enda sammanträde. Då de fasta arvoden alltmer börjar likna månadslöner blir fritidspolitikerna beroende av dem och helt ointresserade av att lämna minsta uppdrag.

Jag vill uttrycka min fulla förståelse för att de förtroendevalda i Stockholm anpassar sig till de system som finns. Det är inte konstigt att en politiskt engagerad pensionär med stor fritid drygar ut pensionen genom att sitta på 15-20 uppdrag. Vi måste dock ställa oss frågan hur detta påverkar demokratin i Stockholm.

<<<kk52>>>

LIX values	Band1, %	FBand2, %	FBands 3-8, %	FB9, %	LFP score	Lex. Density	Lex. Variation	LexLix
29	71.9	6.9	9.3	11.8	39.9	55	41	34.5

<@@@>

@Kropp mot kropp.

@Framsida mot baksida.

@En andedräkt i nacken; fadd, förbrukad, men ännu fläktande. Utbuktning i inbuktning.

@Luften var på väg att ta slut. Det kvalmigt, kvalsterbärande sipprade in i dess ställe. Ner i lungorna trängde det, och upp igen så fort de snörptes samman. Tanne hostade till och kippade andlöst efter mer.

@Ständigt blev hon vidrörd; händer, axlar, lår. Ständigt var hon inom räckhåll.

@Någon stod i mittgången, tätt in på henne, och körde armbågen i hennes bakhuvud så fort bussen saktade in. Hon böjde sig framåt, torkade sig med skjortärmen och flyttade längre in utan att titta.

@Illa luktade han också; svett och - var det damparfym? En klackring i guld stötte mot hennes lillfinger.

@Galonen klubbade fast vid henne. Klitsch, klitsch, klitsch, sa den när den särades från låren. Hon reste sig, drog försiktigt i kjolfällan och satte sig igen. Hon tog spjärn med tåspetsarna och parerade rörelserna för att inte komma för nära mannen till vänster, vid fönsterplatsen.

@Hon sneglade på honom; han var tjock nog att ta även en del av hennes sits i anspråk och hade två stora matkassar i knäet. Eftersom han höll i dem och inte i sätet, pressades han framåt vid häftiga inbromsningar och trycktes bak när de åter fick upp farten. Vid en tvär vänstersväng kastades han handlöst åt sidan och pressade sin nakna arm mot hennes.

@Inte länge, för hon var beredd och drog den till sig.

@Men ändå.

@Hud mot hud. Det var just det hon avskydde med bussturer. Vem som helst, utan ögonkontakt.

@För henne var beröring mycket allvarligare än så. Den förtjänas, slumpas inte bort till den som råkar sitta närmast. Den väljs och har ett fast pris.

@Hon pressade handen mot strupen och dämpade hostreflexen. Mittdörrarna öppnades; äntligen lite luft. Hon harklade sig, behärskat, och drog in det jolmiga genom näsan.

@Mannen som stått närmast plöjde sig igenom folkmassan och hoppade av. En annan tog snabbt och urskiljningslöst hans plats; som en brunstig hanhund på väg mot första bästa människoben, som en....

@Skulle sanningen fram, hade inte heller hon varit särskilt nogräknad. Inte då i alla fall, förut. Hon kväljdes när hon tänkte på vad hon hållit till godo med. Småtafs, som det här; ljumma armhålur och blomsterspråk. Men hon hade låtit sig väljas och alltid, alltid, sett dem i ögonen.

@ "Tobak för halva priset", ropade en tilltufsad man, som just klivit på utanför Högsolan.

@Luften blev underligt frisk, som om ingen vågade andas. Mannen stoppade biljetten och ett fläckigt kuvert i bakfickan och lutade sig ostadigt mot förarbåset.

@ "Inte? Era jävlar!"

@Blickarna sänktes, fulla av indignation. De stående lämnade plats för honom när han vinglande gick och satte sig.

@Längst bak.

@Så klart.

@Trängseln tilltog för varje hållplats de stannade vid. Det var alltid samma mönster; pensionärer närmast chauffören, medelålders av båda könen så nära utgången som möjligt, och så fyllon och tonårskillar där bak. De parfymrade "damerna" blockerade så gott som alltid den lediga innerplatsen genom att sätta sig närmast mittgången och vägra flytta in. Barn i skolåldern valde den fällbara britsen i mitten, såvitt ingen med barnvagn hunnit före.

@Och så fanns det ju Arne Radio, en alldeles egen kategori, som ständigt åkte runt med sin transistor och tröttade ut chaufförerna.

Appendix 5. Swedish consonant clusters

Initial consonants used for the generation of Swedish potential words. Vowels they can combine with are provided only when there are any restrictions on their combinability:

Initial consonant clusters	Vowels	Initial consonant clusters	Vowels	Initial consonant clusters	Vowels	Initial consonant clusters	Vowels
- (zero consonant)		r		tv	a,e,i,ä,å	mj	ä,u
p		l		tr		fn	a,o,i,u,y,å
t		j		kn		fj	a,o,u,å
k		sp		kv	a,e,i,ä,å	fl	
b		sk		kl		fr	
d		st		bj	ä,u	vr	a,e,i,ä,å
g		sm		bl		spl	a,e,i
m		sn		br		spj	ä,u
n		sv	a,e,i,ä,å	dv	ä, a	spr	a,e,i,ä,å
f		sl		dr		str	
s		pj	a,o,u,å	gn		skr	
h		pl		gl		skv	a,i,å
v		pr		gr		sj	a,o,u,å

Final consonants:

p	v	gg	ld	mb	ng	rd	rt	msk
t	r	mm	lf	mp	ngd	rf	rv	nsk
k	l	ss	lk	md	mt	rg	sk	nst
b	j	ll	lm	mn	nt	rk	sm	ngst
d	pp	ft	ln	ns	ngt	rl	sp	psk
g	tt	gd	lp	nch	nkt	rm	st	rsk
m	ck	ks	ls	nk	ps	rn	tm	rst
n	bb	kt	lt	ns	pt	rp	ts	
f	dd	lb	lv	nd	rb	rs	lsk	

Suffixes:

are	ande	het	ion	el	an	lig	mässig	tiv	era
ing	ende	skap	tion	ism	er	sam	aktig	ant	isera
ning		dom	en	else	ig	bar	isk	na	a

Prefixes:

o	miss	för	jätte	ad	veder	gen	er
van		be	an	om	und	fort	de

Appendix 6. Implementation of SCORVEX Modules – some facts

Implementation of C-test Items

The vocabulary generator (i.e. the authoring tool) consists of several modules. Each module is implemented in Java 5.0 as a frame. In this module the following 6 classes are used (see **Error! Reference source not found.9**):

1. GapCloze Frame
2. GapClozeStructure
3. Reader
4. AutoMarkUp
5. TextSelector
6. Abbreviation

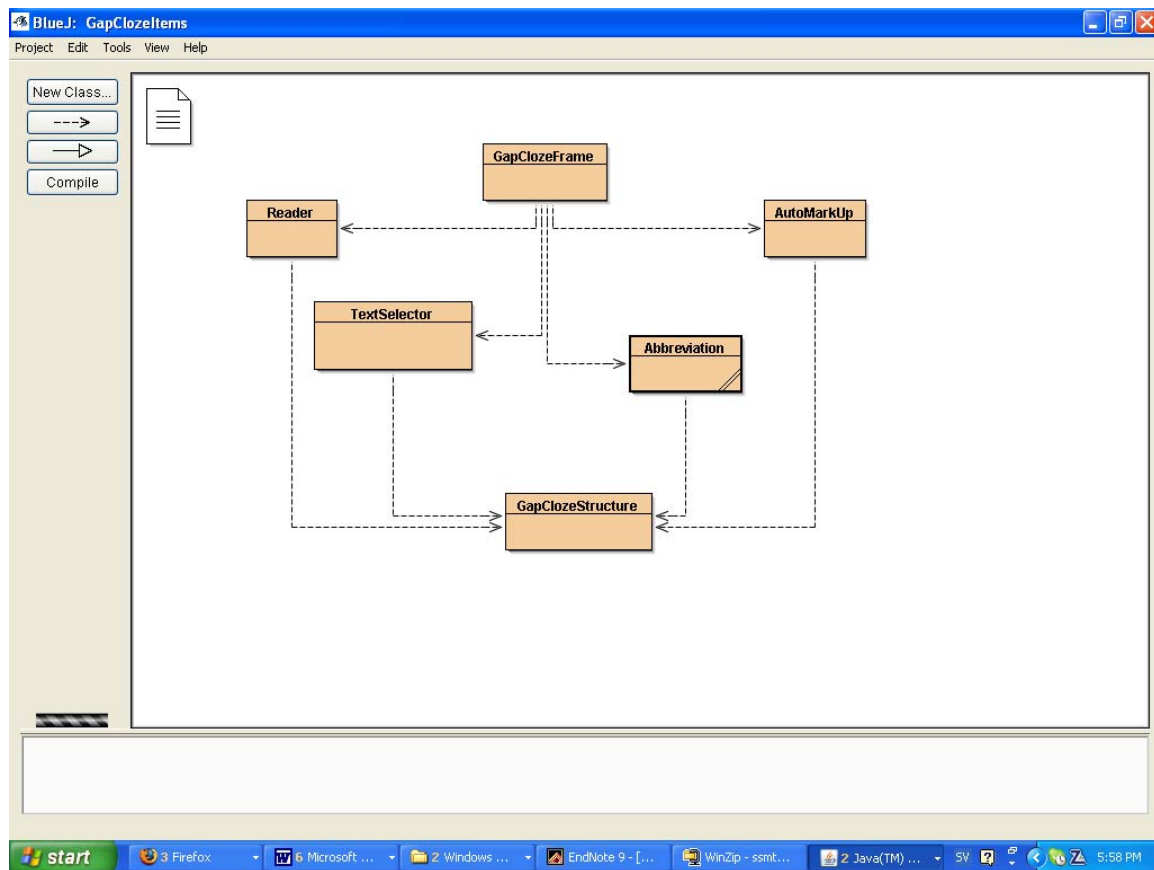


Figure 16. UML-scheme for the C-test Module

Text file archives:

There are a number of files in text format that are extensively used for generation of C-test items, as well as Multiple-Choice items and Wordbank items:

- frequency bands (8 text files, each containing approximately 1000 lemmas);
- lists of SUC files streamed into 6 LexLIX levels, i.e. 6 text files for each LexLIX level containing names of files of a corresponding level;
- archive with SUC texts (500 files);

- archive with SUC sentence references;

A number of classes are reused in several modules. They are: GapClozeStructure, Text Selector, AutoMarkUp and Reader. Classes AutoMarkup and Reader, have slightly different methods that take care of tasks that are specific for each module. Unique classes for the gap cloze module are GapClozeFrame and Abbreviations. In GapClozeStructure there are specific methods for combining all collected information into c-test items.

Depending upon what type of c-tests the user has chosen, GapClozeFrame selects a sequence of methods and commands and calls one method after another. Information from each class and method is stored into the same structures (comes with the class GapClozeStructure – a sort of archivist for all the processes taking place in this module) that follow from method to method.

1. Class TextSelector

Class TextSelector contains 4 methods:

```
public void selectText(int level, int length()
private void selectText()
private void addFiles(int level)
public String textToString()
```

Input to TextSelector should contain student level (1-4) and the length of the extract in number of words to be used for exercise generation. By default 150-word long passages are extracted. All extracted information is stored into ArrayLists – text extract word by word, tags and lemmas for each text.

Method textToString() makes a string presentation of the text that can be printed where the user wants, for example into the user interface window.

2. GapClozeStructure

The information collected into the enumerated below structures during the program run is combined into different types of exercises:

```
ArrayList<String> text, markedWords, abbreviations, baseForms, la_list, letters, lemmas, tags, cTest;
ArrayList<String> originalList, sentenceStartList, sentenceEndList, targetWordsList;
String[] mch_keys, la_keys;
ArrayList<Integer> indices;
TreeMap<String, ArrayList<String>> mch_distractors;
ArrayList<String> specifiedTags;
ArrayList<Integer> freq, currNr;
```

C-test:

```
public String ExeToString()
public String FacitToString()
```

Multiple Choice Exercise, text-based:

```
public String MultiExeToString()
public String MultiFacitToString()
```

Multiple Choice Exercise, sentence-based:

```
public String MultiSentenceExeToString()
public String MultiSentenceFacitToString()
```

Wordbank Exercise:

```
public String ListAnswExeToString()
public String ListAnswWordsToString()
public String ListAnswFacitToString()
```

Printing correct sentences into the frame window:

```
public String ListSentences()
```

3. Class Reader

Class Reader is designed to read the user input from the frame window and fill the appropriate structures with the elements necessary for successful creation of exercises. There are three methods used for reading different types of user input:

```
public void read(Scanner sc)
public void shortenWords()
public void addConsonants()
public void addVowels()
```

Results of these methods are stored into structures: text, marked words, indices and abbreviations.

4. Class AutoMarkUp

Class AutoMarkUp is responsible for automatic selection of words for training. When those have been marked manually, class Reader handles them. Following methods are used for this:

```
public GapClozeStructure markup(GapClozeStructure gcs, int freqBand, ArrayList<String> wordclasses)
public GapClozeStructure markUpAutoText(GapClozeStructure gcs, int freqBand, ArrayList<String>
wordclasses)
private GapClozeStructure selectGaps(GapClozeStructure gcs, ArrayList<String> matches,
ArraList<Integer> tempIndices)
```

Method markup identifies all words of a specified frequency band or of a specified wordclass in a text pasted by the user; method markUpAutoText handles the same but from the SUC text. The difference between the two methods is the way words of a certain wordclass are searched. If the text is automatically selected, all tags are already available and the procedure consists in looking up an ArrayList with tags and selecting words at corresponding indices, whereas with manually pasted text the text needs to first be matched against FLs to identify words of the necessary FL or of the target wordclass. Method selectGaps selects reasonable amount of gaps following at a reasonable distance from each other. Result of these methods is stored into gcs structures: markedWords, indices, specifiedTags, currNr and freq.

Implementation of Multiple-Choice Items

Implementation is made in Java Frame. Seven classes are used for this module (see **Error! Reference source not found.10**):

1. MultipleChoice Frame
2. GapClozeStructure
3. Reader
4. AutoMarkUp
5. Distractors
6. TextSelector
7. SentenceSelector

For information on classes used in all of the modules - GapClozeStructure, Text Selector, AutoMarkup and Reader – see the section on C-tests. I can only add that in AutoMarkUp class in Wordbank Items and in MultipleChoice Items a new method for automatically selecting words of a defined wordclass is added. SentenceSelector is a class shared by both Wordbank items and Multiple Choice Items. It takes care of both automatic selection of target words from FBs and/or wordclasses and handling reader-fed target words for training. Specific classes for multiple-choice module are MultipleChoiceFrame and Distractors. I will describe here classes Sentence Selector and Distractors.

Depending upon what type of multiple choice exercise the user has chosen, MultipleChoiceFrame selects a sequence of methods and commands and calls one method after another. Information from each class and

method is stored into the same structures (comes with the class GapClozeStructure – a sort of archivist for all the processes taking place in this module) that follow from method to method.

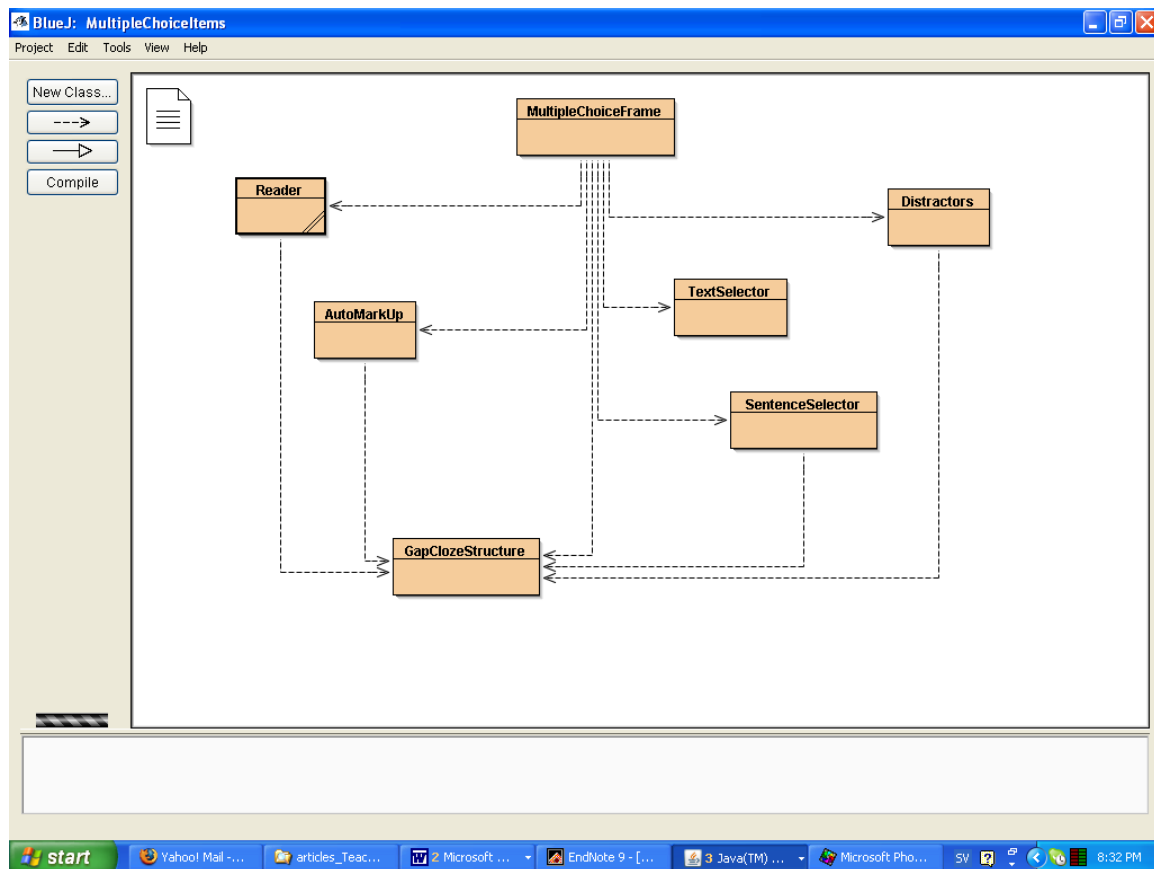


Figure 17. UML-scheme for the Multiple Choice Module

1. Class SentenceSelector

This class is designed to select sentences containing target words and of a required learner level.

SentenceSelector class contains the following methods:

- public void select(GapClozeStructure gcs, int level)
- private void select(String lemma, int level)
- private randomSentence(String lemma)
- private void lookUp(String filename, String sentenceId, String lemma)

Method select(GapClozeStructure gcs, int level) makes use of the ArrayList<String> from gcs with lemmas and their POS-tags that have been either scanned from the user interface or automatically selected from the specified frequency band. Then, from this structure, one lemma is extracted and sent to the method select(String lemma, int level). In this method lemma is used to create a file name and a corresponding file is consulted for sentence-ids of the specified level. If none are found, the next level is searched for. A list of potentially possible sentence ids is created.

In randomSentences(...) one sentence-id is randomly selected from a list of sentence-ids, a file name where the target word is used is extracted from the relevant sentence-id and the information is sent further on to the method lookUp(...), where the sentence itself is collected, split into parts and stored into four ArrayLists: sentenceStartList, sentenceEndList, markedWords, baseForms. Each of these lists is stored in the GapClozeStructure and made use of in the next steps in the program.

2. Class Distractors

Class distractors, as the name suggests, is used to find and organize distractors for multiple choice items. There are six methods, two of them are auxiliary.

- public void findDistractors(GapClozeStructure) makes use of the following structures: baseForms, currNr and markedWords and has as its output freq, updated currNr and specifiedTags
- private boolean find_lemmas(gapClozeStructure gcs, String lemma, String word) is an auxiliary method that is called from findDistractors and helps create the correct output for findDistractors
- public void collectDistractors(GapClozeStructure gcs) uses freq, specifiedTags and markedWords and produces mch_distractors with the help of an auxiliary method
- private ArrayList<String> collectDistractors(GapClozeStructure gcs, String tag, ArrayList<String> distractors)
- public void orderMultChoices(GapClozeStructure gcs) takes care of the distractors, orders them and binds them to the correct gaps in the text or sentences.
- Finally, there is method public void addFiles() that adds necessary files for scanning for search of distractors

At present, as has been mentioned before, a number of classes still need to be implemented, i.e. a class that should take care of saving the existing exercise in QTI format or in text format;

The program relies heavily on an archive with text files:

- frequency bands (8 text files, each consisting of approximately 1000 lemmas);
- lists of SUC files streamed into 6 LexLIX levels, i.e. 6 text files for each LexLIX level containing names of files of a corresponding level;
- archive with SUC texts (500 files);
- database with SUC lemmas – 69,200 files; each such file contains sentence ids where lemmas are used with the corresponding LexLIX level of the text where each sentence comes from.

Implementation of Word Bank Items

Implementation is made in Java Frame. Seven classes are used for this module (see Figure 181):

1. MultipleChoice Frame
2. GapClozeStructure
3. Reader
4. AutoMarkUp
5. Listed Answers
6. TextSelector
7. SentenceSelector

Here, again, a number of classes that are reused in several modules are described. They are: GapClozeStructure, Text Selector, SentenceSelector, AutoMarkup and Reader. Specific classes for word bank items module are WordBankFrame and ListedAnswers. WordBankFrame is responsible that the right order of commands follows each button click, information from all methods is stored into GapClozeStructure. Below follows the description of the class Listed Answers which is unique for this module.

ListedAnswers

Class ListedAnswers, as the name suggests, is used to find and organize answers for gaps. There are six methods, two of them are auxiliary.

- public void orderAnswers(GapClozeStructure) makes use of markedWords and has as its output la_keys and la_list.
- public void orderSentenceAnswers(GapClozeStructure) makes use of markedWords and has as its output la_keys, la_list and targetWordsList.

- private boolean orderWords(String word gapClozeStructure gcs,) is an auxiliary method that is called from orderSentenceAnswers and from orderAnswers and organizers answers in an alphabetically ordered list.

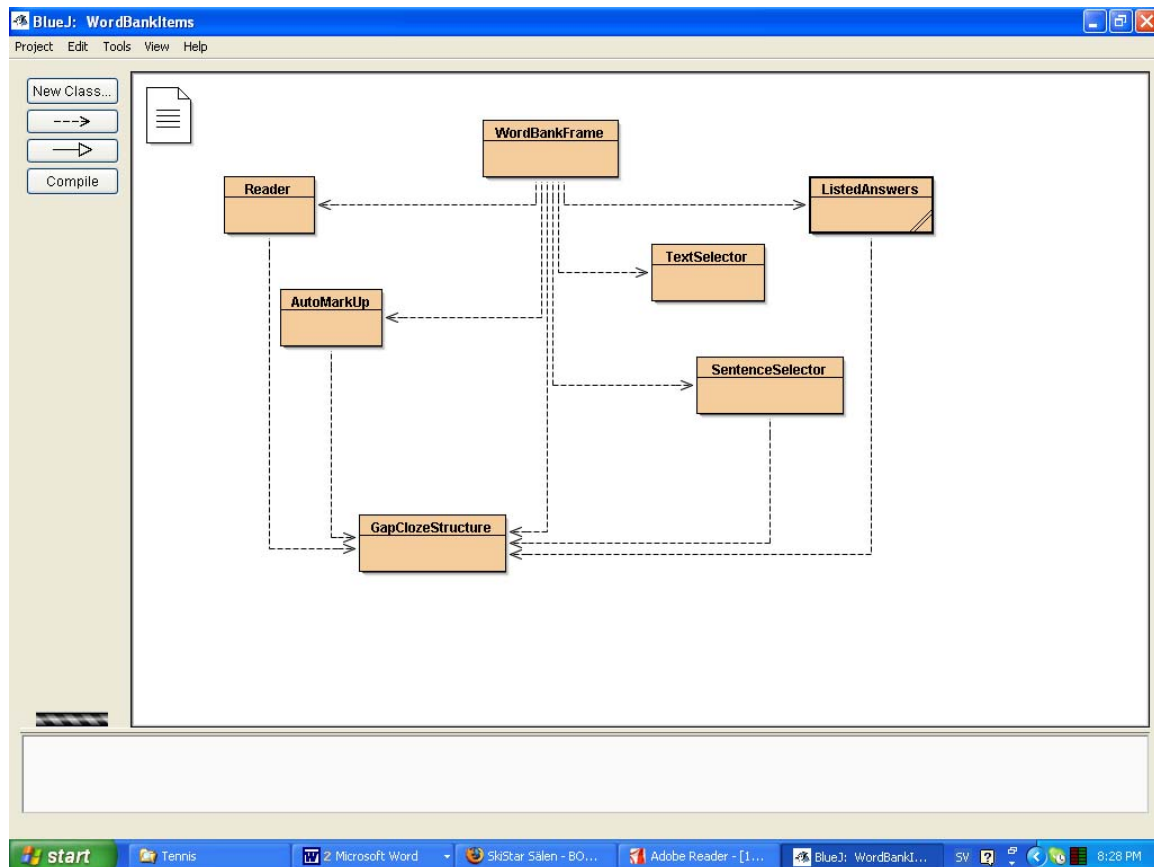


Figure 18. UML-scheme for Word Bank Items Module

The same archive of frequency lists, SUC texts, SUC lemmas and LexLIX levels is used in this module as in the two above described modules.

Implementation of Swedish Vocabulary Size Test

The Swedish Total Vocabulary Test is implemented in Java Frame. The following UML scheme shows the classes and their interrelation in the program (see Figure 12).

There are six operational classes and 4 auxiliary classes that have been used prior to implementation of the program to secure necessary reference files used in program.

The six operational classes are the following:

1. LevelsFrame – the class that takes care of the each pressed button in the user interface if followed by some actions. It is the “main” class in the program.
2. TestStructure contains structures that track information from class to class. There is only one method:
public String toString()
3. TestGenerator is the class that calls all other classes when generation of a test initiated. It contains only one method:

public TestStructure makeTest(int frequencyBand) that calls classes RandomWords and NonsenseWords and fills the two most important structures in TestStructure: ArrayList<String> words and ArrayList<Boolean> values

4. NonsenseWords is the class that coins potential Swedish words of a desirable length (counted in syllables). The following methods make it possible:

```
public ArrayList<String> makePotentialWords(int freqBand)
private ArrayList<String> makeWords(int syllables, int number)
```

The next five methods create lists of vowels, consonant clusters, suffixes and prefixes typical in Swedish for random combining in potential words:

```
public void addInitialConsonants()
public void addVowels()
public void addFinalConsonants()
public void addSuffixes()
public void addPrefixes ()
```

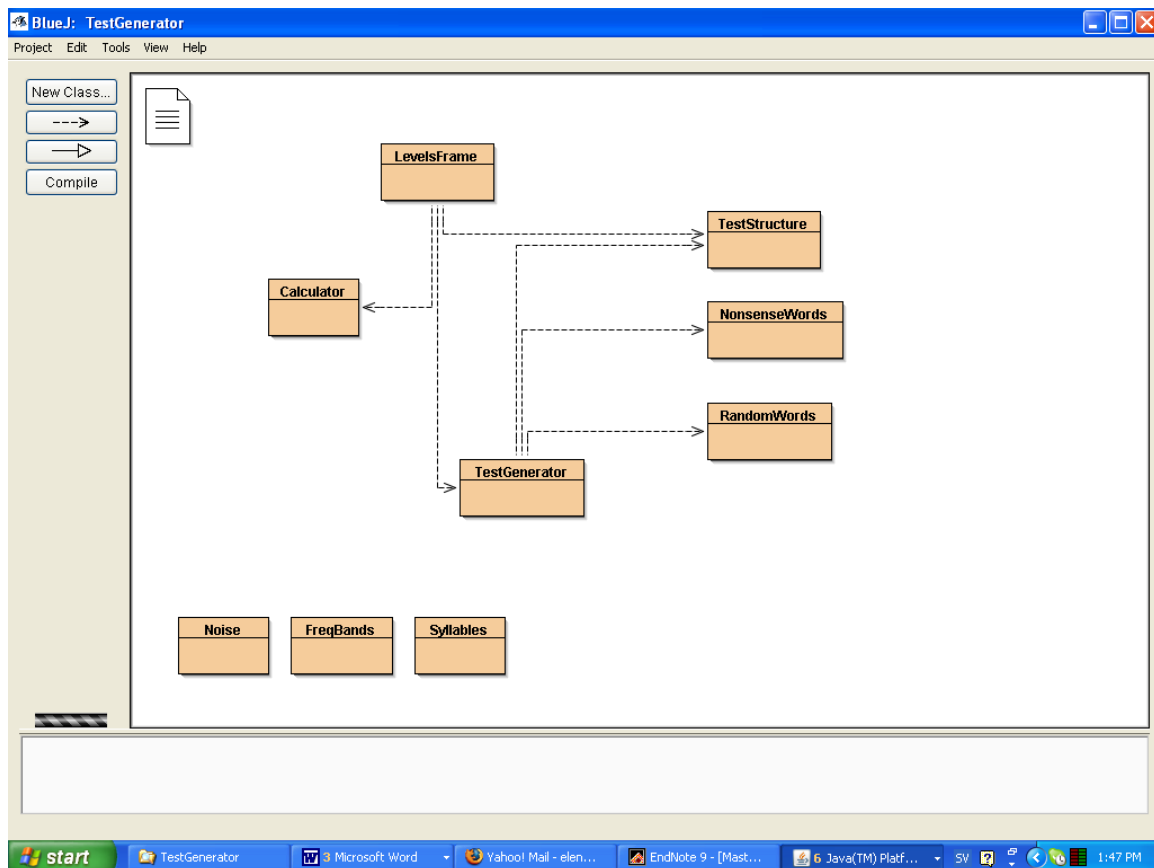


Figure 19. UML scheme for the module on Swedish Vocabulary Size Test

5. RandomWords class selects randomly existing words from a specified frequency band. The following methods make it happen:

```
public ArrayList<String> getRandomWords(int frequencyBand)
private ArrayList<String> getWords()
private String addWord(File file, int num)
private Boolean addInOrder(int num)
```

6. Finally class Calculator counts the right and wrong answers according to a certain algorithm and presents the score on the screen. There is only one method:

```
public void calculate(ArrayList<Double> hits, ArrayList<Double> misses)
```

The three auxiliary classes FreqBands, Syllables and Noise helped to stream base vocabulary pool into 8 frequency bands as well as obtain information about the average amount of syllables per frequency band, and amount of functional words per band.

Frequency Lists are extensively used in generating exercises as well as the lexicon Svenska Ord, that is used to check whether a potential word coined by the program is an existing word with its own entry.