Report on Marie Curie fellowship at BATMULT
(HPMT-CT-2001-00267)

	Mojca Stritar
Period of fellowship:	**15th October 2006 – 15th January 2007**
Project title:	**KUST – Slovene Learner Corpus**

1. Project description

The overall aim of my project is the theoretical foundation of KUST, the Slovene Learner Corpus (SLC). During the Marie Curie fellowship at BATMULT, two major scientific challenges have been faced: the development of a reasonable set of criteria for the collection and selection of learner material to be included in KUST, and the development of an appropriate error tagging system. A preliminary proposal for the selection of material based on a combination of the demographic and contextual sampling has already been worked out, as well as a proposal for the error tagging system tested on a 600-word version of SLC. The actual collection of learner texts has also been started before the beginning of the fellowship.

The main aim of my stay at BATMULT as a Marie Curie Host PhD student was to digitize and tag the material to compile a pilot learner corpus of Slovene based on texts written by learners on different levels of competence and with different first languages. The purpose of the pilot corpus was to check, and if necessary, redefine the criteria for the collection, selection and documentation of learner materials, to develop and test mark-up conventions and the error tagging principles, and finally, to show some possibilities for the use of such corpora for language description, analysis and teaching. The pilot SLC was an important practical part of my PhD research that was still missing at the time. BATMULT thus offered me a unique insight into the Norwegian learner corpus ASK and application of some of its solutions to Slovene language situation. The fact that Paul Meurer at BATMULT speaks Slovene was also extremely helpful to the project.

The expected results of the fellowship included (1) a pilot version of the SLC with a user-friendly interface, (2) an improved definition of the SLC text collection, selection and documentation criteria, (3) improved error tagging conventions and (4) guidelines for possible practical application of the SLC.

2. Activities

2.1. Collection and digitalization of the material

Before the beginning of my BATMULT fellowship, I have collected and mostly digitized the learner material to be included in the pilot SLC. Instead of being well balanced, the design of the corpus was as varied as possible to thus get the chance to test different mark-up and tagging principles. 128 different learner texts have been typed manually with a careful recheck to avoid typing errors.

2.2. Mark-up and tagging

Documents have been encoded in XML following TEI guidelines. The DTD used was basically the same as for ASK. Certain language specific changes have been made regarding the name, but not the content of the document header information. Significant differences, however, are to be found in regards with the error and POS-tagging and will be explained in chapter 3.2.

Document headers have been automatically prepared in MS Access and then imported into Oxygen XML Editor where manual error tags have been added. This software has proven to be an elegant semi-automatic tool and had no problem in processing Slovene characters. Still, the manual work has taken a considerate amount of time, approximately one hour for typing and tagging of each text. Due to adaptation of certain principles during the tagging and also to avoid tagging errors, all the tags have been rechecked.

In questions and problems related to mark-up and tagging, the advice and help of Paul Meurer, researcher at AKSIS, proved to be extremely useful.

2.3. Implementation of SLC material into the ASK interface

The XML-documents have been implemented into the Corpus Workbench tools by Paul Meurer. The SLC web-interface is similar to the one of ASK. SLASK, a miniature one-text Slovene learner corpus, has already been made two years ago by Jana Zemljari Miklav i participating in BATMULT, and the system has been upgraded for pilot SLC.

In comparison to ASK, certain changes have been made to search categories according to the tagging principles. A Slovene translation of the interface has been prepared as the current Norwegian one is not suitable for Slovene-speaking users (at the moment the translation is not operable, but it should be in the near future). Slight design changes have also been applied to the XSL-stylesheet to thus make the secondary errors less emphasized in the output.

2.4. Redefining the error classification

Before and during the actual tagging, several attempts of error classification have been made. They have been tested on actual texts and discussed in live and electronic communication with Paul Meurer, Kari Tenfjord (professor at the Nordisk Institut at the University of Bergen and head of the ASK project), Marko Stabej (professor at the Department of Slovene Studies at the University of Ljubljana), Nataša Pirih Svetina, Jana Zemljari Miklav i (researchers at the Centre for Slovene as a Second/Foreign Language at the University of Ljubljana) and Andreja Markovi (Slovene as a FL teacher).

Following Paul Meurer's advice, I have also set up a "parser" which allows for more objective error classification and will be described in chapter 3.2.

Having decided upon the error classification, some texts have been tagged and a basic error tagging manual has been set. Throughout the work, however, the material showed that certain changes needed to be applied in order to make the classification more transparent and consistent. The second reading of the XML-documents helped clarify and objectify the tags.

2.5. Redefining text collection principles

Through specialized literature, discussions with Kari Tenfjord and growing personal experience, the preliminary text collection principles for SLC have been revised and will be described in chapter 3.3.

2.6. Analysis

Although one of the goals of my fellowship was the creation of guidelines for analysis and practical applications of SLC, it hasn't been achieved completely due to the unbalanced design of

the pilot corpus. Nevertheless, the data has been analysed to test the possibilities offered by Corpus Workbench, i.e. concordance and collocation/colligation lists.

2.7. Meetings and other activities

Frequent meetings and electronic cooperation with Paul Meurer have been one of the cores of my BATMULT stay. Extensive discussions with Kari Tenfjord on 24th October and 19th December 2006 gave me a useful insight into the design and collection of ASK.

I have attented the ASK-workshop in Bergen on 6th December 2006 held in Norwegian, and the lecture "Harmonizing Semantic Resources (about mapping WordNet, FrameNet, VerbNet, SUMO)" by Nancy Ide (Vassar College, New York) on 30th November 2006.

Using the resources of the UiB library, e-brary and access to different on-line journals, I have also gathered new theoretical knowledge on learner corpora, error analysis and second language acquisition in general.

Although not directly related to my PhD studies, the 48-lesson beginner Norwegian language course at the Folkeuniversitetet Hordaland has taken up a considerate amount of my time. My Norwegian language competence has increased significantly, so I am able to follow certain lectures in Norwegian and can also use the pilot SLC corpus tools more in depth.

3. List of results

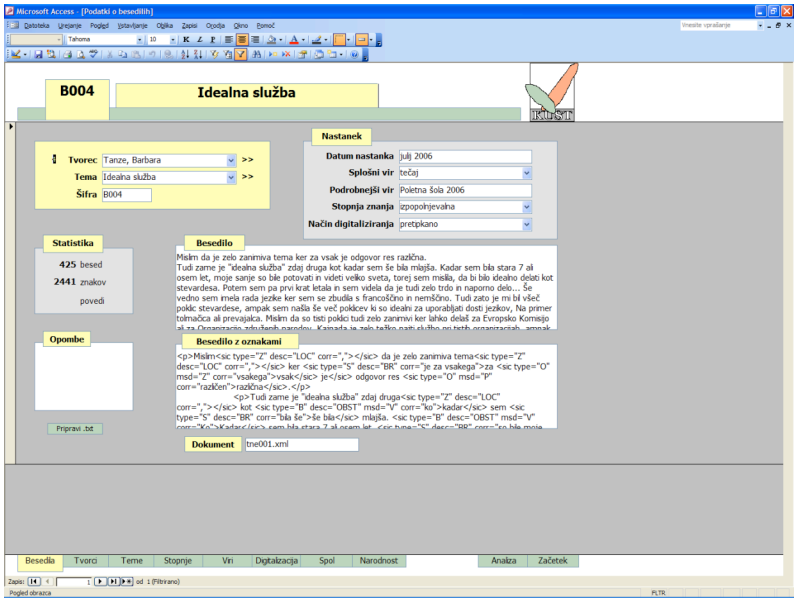
3.1. Pilot Slovene Learner Corpus

The most important result of my BATMULT fellowship is the pilot version of SLC (http://decentius.aksis.uib.no/corpus/norwegian-corpus.html?corpus=KUST). Apart from the fact that the user interface is partly in Norwegian, it is completely searchable and can thus be used to demonstrate the possibilities offered by SLC. Corpus tools allow searching according to different criteria and also advanced CQI-search. Concordance lists and collocations with different layouts can be produced and downloaded for further manual analysis. There is also the possibility of viewing whole texts with our without tagging.

Pilot KUST has 34,873 words in 128 texts written by 119 learners with 18 different first languages (Bosnian, Bulgarian, Chinese, Croatian, English, French, German, Hungarian, Italian, Macedonian, Polish, Romanian, Russian, Serbian, Slovakian, Spanish, Thai and Ukrainian). Texts are from the Slovene language exam for foreigners at the intermediate and higher level in 2001 (91.8%) and from different Slovene language courses in 2005 and 2006 (8.2%). Learners were at a higher beginner (0.09%), intermediate (2.29%) and advanced (96.8%) level of language competence. Texts are mostly argumentative essays on different current topics, but there are also some diaries and official complaints. As expected, the latter have proven to be less suitable for such corpora.

Unmarked texts are collected in a MS Access database created for this purpose. Its main function is transparent document storage and browsing, but it can also provide information on subcorpora sizes or automatically prepare XML-files for each document.

Figure 1: Interface of the SLC database in MS Access



3.2. Error classification

The two-level error categories used in pilot SLC are shortly listed in Table 1.

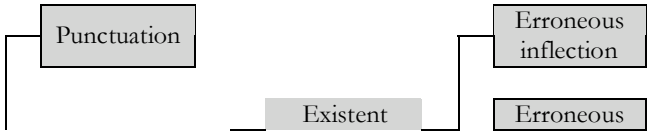
Table 1: Error categories in SLC

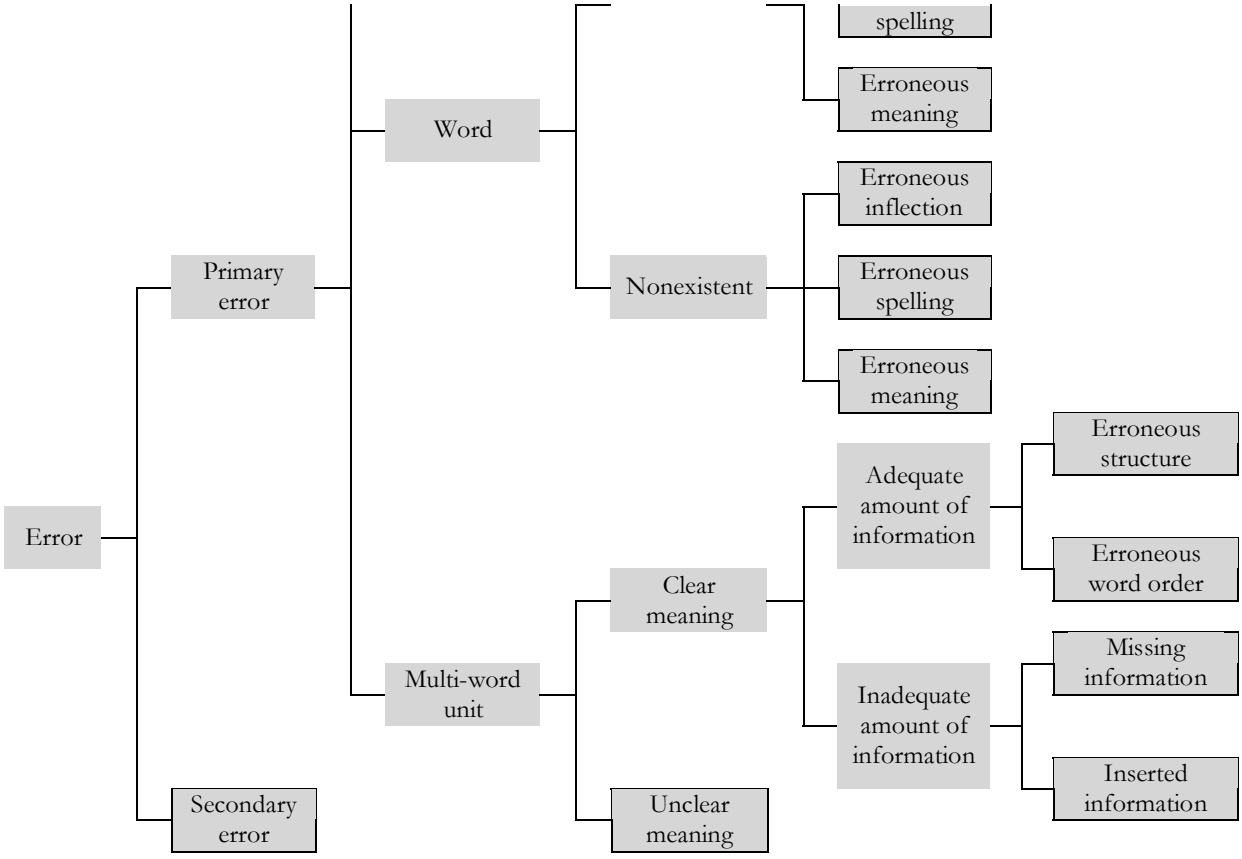
Level 1	Level 2
Orthography	Spelling
	Word division/fusion
	Capitalization
	Punctuation
	Secondary error
Lexical errors	Existent word
	Nonexistent word
	Secondary error
Morphological errors	/
	Secondary error
Syntactical errors	Erroneous structure
	Unclear meaning
	Word order
	Omission of word/phrase
	Insertion of word/phrase
	Secondary error

One-word errors also have a manually assigned POS-tag. Following the Slovene (computational) linguistic tradition these are noun, verb, adjective, adverb, pronoun, numeral, conjunction, preposition, particle, interjection and abbreviation. Corrected forms are assigned to each error, unless it was impossible to infer the intended meaning.

Manual parser with which errors are divided into categories is shown in Figure 1.

Figure 2: Error parser for pilot KUST





An error-tagging manual with examples, problems and solutions has been written and is one of the core chapters of my PhD thesis.

3.3. Text collection principles

SLC collection principles have been revised; to make the data more objective, comparable and easier to collect, all texts should be gathered at Slovene language exams at the intermediate and advanced level (B2 and C1 according to CEFR). Only where there aren't enough learners from a certain language group taking the exam (for instance the English or German speaking), language courses participants could take a similar test specifically for corpus purposes.

First languages of learners represent the biggest and most frequent groups that are also most relevant for language researchers, teachers and testers. SLC should consist of 7 balanced subcorpora and of a bigger subcorpus made of all other languages. The size of the balanced parts is given in Table 2.

Table 2: Proposed structure of KUST

First language	Level of competence	No. of words	No. of texts
Croatian, Serbian, Bosnian	B2	18,000	150
	C1	22,500	150
Macedonian	B2	18,000	150
	C1	22,500	150
German	B2	18,000	150
	C1	22,500	150
English	B2	18,000	150
	C1	22,500	150
Spanish	B2	18,000	150
	C1	22,500	150

First language	Level of competence	No. of words	No. of texts
Italian	B2	18,000	150
	C1	22,500	150
Russian	B2	18,000	150
	C1	22,500	150

A chapter on more detailed description of SLC structure is as a central part of my PhD thesis.

3.4. Error analysis

A simple count of most frequent errors has been done according to different criteria, for instance first language or task-related criteria. The relatively regular distribution of first-level categories (orthographic, lexical, morphological and syntactical errors) has proven they have been chosen well. Table 3 shows frequency of level 1 and 2 error categories; further analysis is beyond the means and scope of current research.

Table 3: Frequency of level 1 and 2 error categories in pilot SLC

	Level 1	Level 2	No. of words
1	Syntax	Word order	1358
2	Orthography	Punctuation	1284
3	Morphology	/	859
4	Lexical	Existent word	735
5	Orthography	Spelling	347
6	Lexical	Nonexistent word	309
7	Syntax	Structure	198
8	Morphology	Secondary error	196
9	Syntax	Insertion	151
10	Syntax	Omission	134
11	Orthography	Word fusion/division	118
12	Syntax	Unclear meaning	113
13	Syntax	Secondary error	76
14	Orthography	Capitalization	34
15	Orthography	Secondary error	14

3.5. Presentation

I gave a presentation of my project at the Bergen Friday Linguistics Seminar at the Department of Linguistic Studies at University of Bergen on 5th January 2007. The feedback turned out to be very constructive and proposals for future cooperation between universities in Bergen and Ljubljana in the field of learner corpora have been made.