An Esperanto spoken language corpus: A proposal for a pilot study

Christer Lörnemark

May 5, 2006

1 Introduction

This project plan is about creating a corpus of spoken Esperanto of 100,000 words, with the aim of enlarging the corpus later, preferably up to and over 1 000 000 words. The corpus linguistic studies on Esperanto are very few ([8] for written Esperanto, [15] for spoken Esperanto). Recording, transcribing, and assembling a corpus of spoken Esperanto may necessitate other routines, other transcription conventions, maybe even other methods, than those used e.g. in building the Gothenburg Spoken Language Corpus (GSLC), one of the biggest corpora for spoken Swedish (1,4 million words), and other spoken language corpora of comparable size for other ethnic languages. (The GSLC will be used as a model for the Esperanto spoken language corpus with regard to transcription standard and the activity-based construction of the corpus.) Therefore, a pilot study has been deemed necessary.

2 Some uses of a spoken Esperanto corpus

There are several uses for this corpus, as well as research questions to ask when exploring it. These are a few examples:

- Making frequency lists for roots, words, and collocations. These lists could be used for comparisons with the written Esperanto corpus, for various Esperanto language courses, and for studies on Esperanto phraseology (e.g. for comparisons with [7]).
- Getting to know the most frequent differences between the written language norm and spoken Esperanto as it actually is in the transcriptions.
- The use of Esperanto in different activities: E.g., how is the Esperanto of the meeting of a local Esperanto club different from the Esperanto of an interview of Pola Radio, or from coffee break talk at the Centra Oficejo in Rotterdam?

- The influence of the native language on an individual's use of Esperanto (lexical choices, grammar, pronunciation).
- Investigating non-verbal behaviour - and the interplay between verbal and non-verbal behaviour - among Esperanto speakers.
- Esperanto's status as a language has been and is debated by some linguists, and also by some non-linguists. In both cases there is a lack of knowledge about Esperanto and its daily use in the world. The existence of a spoken Esperanto corpus may help spread the word among linguists that Esperanto is almost like any other language, expressive and in use, and that it is worth serious linguistic research.

(Only the first item, frequency lists, is covered in this proposal. The other items are for projects in the future.)

3 Activities represented in the corpus

Spoken Esperanto will be recorded from a number of activities (see next section). It is crucial to know what kinds of conversations to include in the corpus, and to have a detailed planning at the outset of the corpus building. That is especially important with a small initial corpus of 100 000 words. I will use two existing spoken language corpora, the GSLC for Swedish, and the spoken language part of the BNC (British National Corpus), in order to see how an Esperanto corpus could be assembled.

3.1 The GSLC

The GSLC for Swedish consists of 25 activities [13]:

Discussion
Retelling Of Article
Interview
Task-Oriented Dialogue
Informal Conversation
Role Play
Trade Fair
Arranged Discussions
Formal Meeting
Consultation
Shop
Dinner
Market
Auction
Factory Conversation

Party
Games & Play
Phone
Travel Agency
Court
Church
Lecture
Hotel
Therapy
Bus Driver-Passenger

Here we will group them into 13 groups according to their overall purpose, in order to see how many of these overall purposes are applicable to spoken Esperanto:

Academic: spoken language in academic more-or-less formal settings: e.g. seminar discussions;

Administrative: meetings of local and state authorities, board meetings, etc.;

Commercial: salesman/customer talk: trade fair, travel agencies, shops, supermarkets, etc.;

Health care: patient-doctor consultations, optician-client interaction;

Jurisprudence: court proceedings

Mass medial: radio programmes, TV programmes

Play: games, role play

Political: political debate, parliamentary hearings

Religious: church service

Research: various linguistic experiments

Socializing: informal conversations among family members, among friends, among colleagues

Teaching: lectures

These overall categories are not exhaustive for all activities that may be pursued by means of spoken language. Proposals for new overall purpose categories are most welcome. They are not meant to be philosophical categories that would cover all spoken human activities, merely a way of

categorizing the transcriptions of the GSLC. The GSLC has not been assembled according to these broad categories, but from a subset of the 25 activities in their list, which has expanded gradually to the present list.[1]

3.2 The BNC

We could also look at the spoken language part of the British National Corpus - 10 million words of British English. [4] This corpus was assembled in two parts, using two different categorizations: demographic vs context-governed. 124 volunteers participated in the demographic part; they were from 38 different locations across the UK and were chosen so that there would be an equal number of men vs women, and approximately equal numbers from each age group and each social group. The context-governed part used four broad categories, called "social context", so that one would have roughly equal amounts of words from each of the four categories:

> **Educational** and **informative** events, such as lectures, news broadcasts, classroom discussion, tutorials.
>
> **Business** events such as sales demonstrations, trades union meetings, consultations, interviews.
>
> **Institutional** and **public** events, such as sermons, political speeches, council meetings, parliamentary proceedings.
>
> **Leisure** events, such as sports commentaries, after-dinner speeches, club meetings, radio phone-ins. [4]

(There is also a small part called Unclassified.)

So the two categorizations are different; one could devote a whole article comparing and contrasting the two corpora (and possibly other corpora as well). For the Esperanto corpus the important thing here is to look at the activities that are possible and relatively frequent among Esperanto-speakers.

3.2.1 Possible activities for a big Esperanto spoken language corpus

At first, we may look at a number of activities that are pertinent for a big corpus, say 1 million words. Then, in the next subsection, we will look at the possible contents of a small corpus of 100 000 words.

The Esperanto speakers differ from speakers of most ethnic languages with regard to activities: there is no Esperanto speaking nation state, or

[1] In practise, there are overlaps between those categories, e.g. TV programmes (mass medial) that are made for didactical purposes (Teaching). What is pertinent here is the overall purpose of the conversation itself, not secondary uses of the recording (conversations that are made for one overall purpose and whose recordings happen to be used for another overall purpose).

even ethnic group, and therefore the range of activities in spoken Esperanto is probably smaller than in an ethnic language that is in use today. One could start with a list of conversation types that are likely to be found within the Esperanto micro society:[2]

Academic: linguistic discussion about Esperanto; lectures (within Kongresa Universitato or the Akademio Internacia de Sciencoj San Marino); seminar discussions (within the AIS)

Administrative:

1. organizational meetings within the Esperanto movement: boards, committees, work groups, annual meetings, Komitato meetings[3];
2. work related conversation: talk between professional Esperantists (office workers) at UEA's head office. Here Esperanto is the working everyday language;
3. conversation between members of e.g. an Esperanto club while carrying out some practical task (e.g. mailing the club periodical, manning a stand at an exhibition etc.);[4]

Commercial: a stand for the selling of souvenirs at an Universala Kongreso; the Excursion desk at the UK

Mass medial: interviews, e.g. in Esperanto language broadcasts (radio and TV)

Play: role play; games

Religious: a sermon in Esperanto (provided that it is not read aloud from a manuscript)

[2]This list is partly based on [9] and on [13] and [4], using the overall purpose categories from a previous subsection.

[3]Universala Esperanto Asocio is the umbrella organization of the international Esperanto movement. UEA has about 20 000 members in 115 countries. The members are organized in 63 national Esperanto organizations; there is also a category of about 3000 individual members. The individual members and the national Esperanto organizations are represented in Komitato, which has its annual meeting during UEA's yearly world congress.

[4]The degree of Esperanto use varies between Esperanto clubs. The first language of these Esperantists is mostly used, simply because it comes naturally, while use of Esperanto is most often a conscious decision that is not always reciprocated by other speakers (e.g. by Esperanto beginners, who may prefer always to use their native language (or some other language used in that country) with other members of the club. When there is a foreign Esperantist present, the members strive to use Esperanto as much as possible, so that the foreigner may not feel left out. So a recording of practical task talk in an Esperanto club could be "artificial" in the sense that the participants have to be asked beforehand to avoid using ethnic languages. A small amount of code switching is probably inevitable in some settings and could be the subject of future studies.

Research: retelling of a text

Socializing: conversations for socializing (informal conversations);

Teaching: teaching Esperanto (from beginner level to advanced level, from informal courses to tertiary education, e.g. at the Adam Mickiewicz University in Poznan);

The following overall purpose categories are missing from the above list:

- Health care (does not exist in Esperanto)
- Jurisprudence (no judicial system exists in Esperanto)
- Political: There are wide differences in political outlook among Esperanto speakers, but very few arenas in which to conduct a political debate, since there is no Esperanto state with a parliament. One possible environment is the socialist organization SAT. That could yield discussions/debates among left-wingers. There is no organization for liberals or right-wingers within the Esperanto movement.

(However, there are few obstacles to discuss these topics in Esperanto.)

In contrast to ethnic languages, the Esperanto speech community is largely a second language community.[5] It is therefore pertinent to use methods from the research on second language acquisition. Several such methods were used within the EALA project (Ecology of Adult Language Acquisition), a huge project (sponsored by the European Science Foundation) that was carried out 1982-1988 in four countries. [6] Immigrants were used as informants and studied longitudinally for several years. The researchers used the following methods: sociobiographic interview, role play, informal conversation (with the aim of "warming up" the subject before an experiment), various linguistic experiments, route descriptions, self-confrontation (recording an interaction that was made with one of the other methods, showing it to the informant and asking why he/she used a certain construction), accompanying observation (the informant does his/her errands, e.g. going to an employment exchange office, and the researcher follows him/her and makes a recording). Some of these methods could be used for the Esperanto spoken language corpus as well. Some of these recordings would be a kind of laboratory speech that would not necessarily mirror the spontaneous speech e.g. at Esperantists gatherings. In the corpus one would have both spontaneous speech and speech that would be wholly or in part prepared and prestructured.

It is essential that it may be possible to compare and contrast speech according to these parameters:

[5] Individuals who have learnt Esperanto in their childhood should be included in a big Esperanto spoken language corpus. That would give a chance to investigate differences between their Esperanto and the spoken language of ordinary Esperanto speakers.

- Women vs men
- Face-to-face vs telephone conversations
- Public vs private speech
- To have different interaction types[6] in the corpus:
  - argumentation
  - debate
  - interview
  - monologue
  - quarrel
  - conversation (for those interactions that do not fit into any other interaction type)

But one also has to be pragmatic: there are budget and timeline restrictions, as well as possible difficulties in obtaining recordings from a large number of activities. There are probably a large number of Esperanto recordings floating around in the Esperanto speaking community, and it would be counter-productive not to include some of these in the Esperanto corpus. So this long list should not be taken as a guarantee that everything in it would some day be included into a big spoken language corpus. It is merely a point of departure for assembling a corpus. Also, no corpus could be all-inclusive, so whatever is done to it it will always be just a window into spoken language.

3.3 Possible activities for a small Esperanto spoken language corpus

For a corpus of 100 000 words it is most practical to use a few activities rather than the whole list (or most of it) that was covered in the previous subsection. Here is a proposal for these few activities:

- Administrative: two hours from the annual meeting of the Komitato (a whole four hour session could be recorded, and the remaining hours could be transcribed and added to the corpus later)
- Mass medial: interviews from the Polish Radio
- Socializing: informal conversations at an Esperanto club and at an Universala Kongreso

[6]There is one other interaction type, interrogation, in the GSLC (a parliamentary hearing), but such a thing (not to speak about police interrogations) are very unlikely to occur in Esperanto, so it was left out here.

- Teaching: lectures from Kongresa Universitato

Each of these four categories would occupy about 25 000 words each. The recordings will probably be of different lengths, and e.g. if the whole Komitato session (not to speak of mettings of Subkomitatoj) were to be transcribed at once, most of the small corpus would consist of administrative, follow-the-agenda talk. Therefore, sections of individual recordings could be left untranscribed for this small corpus, and be transcribed and added later.

4 Making recordings

The recordings will be made by one or two individuals. The recordings should be made multimodally, so that verbal and/or non-verbal transcriptions can be made from them. That is especially important for recordings with a high number of speakers, e.g. a Komitato meeting, which has about 50-60 participants; the transcriber has to be able to see who is talking. For conversations with a few participants an audio recording may suffice in many cases. The recordings should be as ecologically valid as possible. [10]

For this small corpus we propose two venues of recording: the Esperanto club of Göteborg, Sweden, and the Universala Kongreso in Florence (2006). In the first case one would obtain Esperanto spoken by Swedish Esperantists, in the second case not only Esperanto speech by Italians, but also by a large number of ethnicities (an Universala Kongreso held in Europe usually has participants from about 60 countries). A congress of that kind is such a good venue for recordings that about 30 hours of recordings could be made (providing material for a larger corpus in the future).

There are many recordings of spoken Esperanto in the world, e.g. recordings made by individuals at Esperantist gatherings, and recordings in the sound archives of Pola Radio, Universala Esperanto-Asocio, and Internacia Esperanto-Muzeo. The only obstacles for including such recordings are their technical quality, the activity being pursued in them (in order to get a balanced corpus), and that we get permission to use them from the copyright holder and the individuals who have been recorded.

All recordings have to be digitized. Using analogue recordings would necessitate cassette recorders (not to speak of older tape recorders for older tape formats), and that should be avoided, as the playing and rewounding of small parts of the tape - during transcription - takes too much time. Also, a digital format is easier to convert to newer storage formats and to newer codecs.

5 Transcription

We propose that the Göteborg Transcription Standard [12] be used for the small corpus. The GTS is a language-independent standard, so small adap-

tations probably have to be made for Esperanto.

In order to ensure the validity of the transcriptions there has to be one transcriber and one checker.

Transcribing the data is very time consuming. One recorded hour may take up to 60 hours in total to transcribe and to check. A recording containing many participants and many stretches of overlapped speech takes longer to transcribe than a recording where 2-3 participants alternate without interrupting one another frequently. A recording of bad technical quality may take longer to transcribe than a high quality recording. It is crucial to maintain a high quality of transcriptions from the very beginning, and to apply the transcription standard consistently across the whole corpus. Taking shortcuts in the beginning, e.g. to tolerate minor errors in the transcriptions in order to build volume and get results quickly, would only lead to higher costs later (where the transcriptions would have to be corrected once again). Therefore, a realistic timeline has to be there from the very beginning.

One hour of recorded speech may contain about 10 000 words. For a 100 000 word corpus one would need 10 hours of recordings. So the transcription and checking could take 600 work hours. Transcribing demands full concentration and is something that can't - or rather, shouldn't - be done on a full time basis; 6 transcription hours is enough for a work day. If we calculate 20 work hours per week for one transcriber, and if we assume that the transcription would take 550 hours and the checking 50 hours we would get 27 1/2 work weeks for the transcriber. The checker could be hired on a per-hour basis.

(If, however, some transcriptions would take less time to transcribe and check than 60 hours per recorded hour that would leave time to transcribe more than 100,000 words.)

The transcription itself would be made with a text editor and a program for analyzing the acoustic signal, so that a small stretch of speech could be marked and played (in earphones) as many times as is needed to transcribe what is said.

6 Analyzing the data

The statistical data may be obtained by experts hired for this purpose. For this small corpus I had a Swedish programmer living in South Korea in mind (Bertil Wennergren). He has been working with the Esperanto written language corpus Tekstaro. If budget restrictions permit one could also - or instead - ask the spoken language research group at the Department of Linguistics, Göteborg University (the creators of the GSLC).

A large number of data can be obtained from the corpus. In this small project we will look at:

- word frequencies
- word root frequencies
- collocation frequencies (for 2, 3, 4, 5, 6, 7, and 8 word collocations)

These data could be made for:

- the whole 100 000 word corpus
- the part Administrative
- the part Mass Medial
- the part Socializing
- the part Teaching

Noone has ever made this kind of statistics for Esperanto transcriptions in GTS, so we would have to be cautious about promising too much within the timeline for this project. The methods that are to be used for this small corpus could later be applied to a much bigger spoken language corpus.

Bertil Wennergren has written a dictionary program system for the Lernu website, and part of that code can be used for semi-automatic morphologic analysis. The dictionary used for this system contains more word roots than the 2002 edition of Plena Ilustrita Vortaro.

7 Storage of the corpus and the recordings

The corpus could be stored e.g. on one or two of the following servers:

- on www.bertilow.com, together with the Tekstaro. (There is room for a corpus of 100 000 words, but not for a bigger corpus.)
- on the Ikso server (owned by Esperantic Studies Foundation). (At present there is room for the corpus itself on Ikso, but not for the recordings. An extra hard drive for Ikso is included in the budget.)
- on a server at the Department of Linguistics, Göteborg University.

One could have the corpus on one of the servers, and a copy on another.

There is an offer from the Department of Linguistics, Göteborg University, to store the corpus on one of the department's servers, and to provide shelf space for the recordings in their tape archive in exchange for collaboration on knowledge and technical equipment. The conditions for this collaboration - and whether there is any additional cost for the ESF - would have to be negotiated between the department and ESF.

The recordings will be in digital format(s), so they don't have to be stored on physical media e.g. CDs and/or DVDs; they could be stored on a server together with the transcriptions. CDs and DVDs could be used for a backup copy to be placed in a sound archive.

8 Access to the corpus

ESF will own the corpus, but it would be accessible to researchers, just like the Tekstaro. A user login system like that used for the Tekstaro could be used for the spoken language corpus as well.

9 Timeline

2006	
April-May	Making recordings with Swedish Esperantists in Göteborg. Asking various Esperanto institutions about permission to use recordings that they may have.
June	Starting the transcription work. Making recordings at the Nitobe seminar in Göteborg.
July	Transcription work. Making recordings at the Universala Kongreso in Florence.
August	Transcription work. Checking transcriptions.
September	Transcription work. Checking transcriptions.
October	Transcription work. Checking transcriptions.
November	Transcription work. Checking transcriptions.
December	Transcription work. Checking transcriptions. Data analysis.
2007	
January	Data analysis. Starting to write the first scientific article on the corpus.

The computations on the transcriptions - in order to get frequency lists etc. - could start in the middle of December. If all goes well, the first statistical results would be ready before Christmas.

Research on this small corpus is not calculated into this budget, because the corpus can be used for so many purposes that these have to be covered by other projects.

10 Budget for an Esperanto corpus of 100 000 words (in USD)

10.1 Personnel costs

Person / personnel category	**Function / Purpose**	**Time/Thing & piece price**	**Sum total**
Lörnemark, Christer	Researcher / Transcriber	6 1/2 months à 3500 (20 hours / week)	22 750
?	Checker	50 hrs à 22	1100
Wennergren, Bertil	Technical personnel	80 hrs á 60	4800
Buffer	Contingencies		6000
Personnel costs in total			**34650**

For residents of Sweden, 65-70 % of the brutto salary goes to the Swedish tax authority (payroll tax + income tax). That applies to Lörnemark. Who will be the checker has not been decided yet. If he/she lives in a country with less taxes (including payroll tax) the salary could be decreased.

In addition, Kaarlo Voionmaa (Dept. of Linguistics, Göteborg University) has said that he is available for advice and help, as far as his work at the department permits.

The buffer is there for unforeseen drawnouts in the corpus building process (regarding transcription, checking, or data analysis), as well as machine failures, loss of equipment (through accident or theft). One should note that there is no insurance for this project. It is hoped that only part of the buffer money would be needed.

10.2 Budget other than personnel costs

Person / personnel category	**Function / Purpose**	**Time / Thing piece price**	**Sum total**
Researcher's travels to Universala Kongreso in Florence 2006	Make recordings	2 travels á 600	1 200

Congress fee	Admission to the congress	1 fee à 300	300
1 portable computer	Data collection and elaboration	1 computer á 1600	1 600
1 external microphone	Recording	1 microphone á 600	600
Portable DVD video camera	Data collection	1 videocamera à 1800	1 800
Other equipment (tapes, batteries etc.)			120
1 IDE hard drive	Storage of recordings	1 hard drive á 250	250
Summed up			**5 870**

References

[1] Jens Allwood. An activity-based approach to pragmatics. In H. Bunt and B. Black, editors, *Abduction, belief and context in dialogue: Studies in computational pragmatics*, pages 47–80. John Benjamins, Amsterdam, 1995. Also published as: Gothenburg Papers in Theoretical Linguistics 76, Dept. of Linguistics, University of Göteborg.

[2] Jens Allwood. Capturing differences between social activities in spoken language. In I. Kenesei and R. M. Harnish, editors, *Perspectives on semantics, pragmatics and discourse*, pages 301–319. John Benjamins, Amsterdam, 2001.

[3] D. Biber, S. Conrad, and R. Reppen. *Corpus linguistics: investing language structure and use*. Cambridge University Press, Cambridge, 1998.

[4] Oxford University British National Corpus. The spoken component of the bnc. `http://www.natcorp.ox.ac.uk/what/spok_design.html`, 2005.

[5] J. Dietze. *Frequenzwörterbuch Esperanto-Deutsch: Die meistgebrauchten Wurzeln der Esperanto-Literatursprache*. Number F96 in Wissenschaftliche Beiträge. Martin-Luther-Universität Halle-Wittenberg, Halle, 1989.

[6] European Science Foundation. *Ecology of Adult Language Acquisition - A Field Manual*, 1982.

[7] Sabine Fiedler. *Esperanta frazeologio*. Universala Esperanto-Asocio, Rotterdam, 2002.

[8] Christopher Gledhill. *The grammar of Esperanto A corpus-based description*. Lincom Europa, Munich, 2000.

[9] Ilona Koutny. Interkultura komunikado en euxropo: la angla kaj esperanto kiel alternativaj komunikiloj. In Ch. Kiselman, editor, *Symposium on Communication Accross Cultural Boundaries*. Kava-Pech, Dobrichovice, 2005.

[10] Christer Lörnemark. Ekologiskt valida videoinspelningar. In Jens Allwood and Elisabeth Ahlsén, editors, *Lingvistisk metod*. Institutionen för lingvistik, Göteborg Universitet, 2002.

[11] Tony McEnery and Andrew Wilson. *Corpus linguistics*. Edinburgh textbooks in empirical linguistics. Edinburgh University Press, Edinburgh, 1996.

[12] Joakim Nivre, Jens Allwood, Leif Grönqvist, Magnus Gunnarsson, Elisabeth Ahlsén, Hans Vappula, Johan Hagman, Staffan Larsson, Sylvana Sofkova, and Cajsa Ottesjö. *Göteborg Transcription Standard*. Dept. of Linguistics, Göteborg University, Göteborg, 6.4 edition, 2004.

[13] Göteborg University Spoken Language Research Group, Dept. of Linguistics. Activity types (under construction). `http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3&SUBPAGE=1`, 2005.

[14] Michael Stubbs. *Text and corpus analysis Computer-assisted studies of language and culture*. Blackwell Publishers Ltd., London, 1996.

[15] Zlatko Tisxljar. Frekvencmorfemaro de parolata esperanto. Zagreb, 1981.

[16] John C. Wells. *Lingvistikaj aspektoj de Esperanto*. Universala Esperanto-Asocio, Rotterdam, 1978.

,