Esperantologio / Esperanto Studies **6** (2013), 109–128

Sources of text for Esperanto corpora

Vilius Normantas

Contents

Abstract. Use of corpora is not a new concept in Esperanto studies. There are two readily available corpora. *Tekstaro de Esperanto* offers a relatively small collection of carefully selected samples of written Esperanto through the entire existence of the language. *VISL* by comparison is much larger and includes more samples, for the most part collected from the Internet. There are certain limitations to both: absence of spoken language samples, absence of manual linguistic annotation, and issues relating to representativeness. The corpora could be improved by harvesting multiple sources of both written and spoken language available on the Internet.[1]

Resumo. *Fontoj de tekstoj por korpusoj de esperanto*
Uzo de korpusoj ne estas nova koncepto en la studoj pri la lingvo esperanto. Ekzistas du prete uzeblaj korpusoj. *Tekstaro de Esperanto* prezentas relative malgrandan kolekton de zorge elektitaj ekzemploj de skribita uzo de esperanto dum la tuta ekzisto de la lingvo. *VISL* kompare estas pli granda kaj enhavas pli da ekzemploj, plejparte kolektitajn sur Interreto. Estas kelkaj mankoj en ambaŭ korpusoj: manko de ekzemploj de parola lingvo, manko de mane kreitaj lingvistikaj prinotoj kaj problemoj rilataj al reprezenteco. La korpusoj povas esti plibonigitaj per kolekto de materialo el multaj skribitaj kaj parolaj fontoj haveblaj sur Interreto.[2]

1. Introduction

The purpose of this paper is twofold. First of all it is an attempt to create a list of sources of Esperanto use suitable for inclusion into corpora. The list is by no means a complete collection of websites in Esperanto—it is rather

[1] *Keywords:* corpus linguistics, corpus design, spoken and written corpora, Esperanto.
[2] *Ŝlosilvortoj:* korpusa lingvoscienco, desegno de korpusoj, parolaj kaj skribaj korpusoj, esperanto.

a collection of examples from the particular registers of language use. Two criteria have been used to judge usefulness of a source:

- How large a sample of language use could be collected from the particular source?
- How rare or difficult to obtain are samples of Esperanto use in that register? Due to this criterion, some relatively small sources are included.

Emphasis has been made on the sources containing written text publicly available on the Internet; however I also touch the topics of printed materials and sources of spoken Esperanto.

The sources could be harvested to create a specialist corpus designed for some particular research question, or to supplement the existing corpora. This brings us to the second intention behind this paper: a call for continuous efforts to improve and expand the existing corpora of Esperanto in order to bring it to the level of the twenty-first-century corpora of other languages (see a discussion about the generations of corpora in Subsection 1.1).

Internet addresses of the websites and individual pages mentioned in the article are provided at the end. To keep the list manageable, I did not include addresses of very well known websites (for example, Facebook or YouTube) or when the address is obvious (for example, Esperanto.com).

1.1. History of Corpora

Over the last few decades corpora of various shapes and sizes became almost indispensable tools in many sub-fields of linguistics and beyond.

Research using large collections of text predates modern corpora by centuries, but it was the appearance and spread of computers which led to significant breakthroughs in the field. Early punched-card-based concordances of the late 1950s (McCarthy & OKeefe 2010:4) could be considered the first examples of corpora use as it is known today. The size of corpora grew dramatically together with advances in computer performance. Elena Tognini Bonelli divides growth of electronic corpora into three generations: (a) 1960–1980, corpora of up to a million words typed manually by keyboard; (b) 1980–2000, availability of scanners and spread of computer typesetting allows to collect corpora of tens of millions of words; (c) 2000– , spread of the Internet in the new millennium makes unprecedented amounts of text easily accessible (Tognini Bonelli 2010:16). Many modern mega-corpora contain from several hundred million to over a billion words.

The present trend toward ever larger corpora is likely to continue, at least as long as the present trend of growth in computer performance permits. However size is certainly not the only relevant measure of corpus usefulness.

A smaller corpus, provided that it is well balanced and representative of the specific area of research, may prove to be at least as useful as a large mega-corpus (Nelson 2010:55). Besides, smaller corpora are generally easier to create, maintain, and work with, which also offers some advantage.

Corpora could be divided into several categories according to their design, contents, and the way they are used. Tognini Bonelli (2010:20–25) describes the following types of corpora:

1. A *sample* corpus aims to represent normal linguistic features at a specific time. When a sample corpus is large enough to represent all the features of a language it may be called a *reference* corpus.
2. When several corpora are designed in the same way, they are said to be *comparable*. The comparison may be geographical, historical or according to topics. A *contrastive* corpus is specifically designed to highlight some difference between two variations of language use.
3. A *special* corpus is a collections of texts selected according to certain criteria other than to represent a general use of language. For example, a corpus of works by a single author.
4. There are two distinct types of corpora that include the time dimension. A *diachronic* corpus consist of snapshots taken at specific intervals in time. In contrast, a *monitor* corpus represents continuous change in language as it occurs.
5. A *bilingual* corpus is designed to compare two languages; *multilingual*, more that two.

The role of a corpus in linguistic research also differs. At the most basic level it may be used merely as a database of examples to support one's ideas derived from other methodologies. However, the principle of total accountability (McEnery & Hardie, Eds. 2012: Section 1.6) encourages researchers not to discard evidence which disproves one's hypothesis. At the theoretical level, the roles of corpora could be divided into corpus-based and corpus-driven approaches (Tognini Bonelli 2010).

1.2. Ethical and legal considerations related to the creation of corpora

Even gathering of publicly available material from the Internet raises a number of ethical and legal issues. Additionally, differences in copyright laws and regulations regarding the use of the Internet between countries may create further problems.

McEnery and Hardie list four approaches which help to deal with copyright issues regarding to content obtained from the Internet (McEnery & Hardie (Eds. 2010, Section 3.2, § 4)):

1. In many cases, text obtained from the web should be treated just as any text collected from other sources. One should contact the copyright holder and ask for a permission to collect and/or redistribute the data. This approach would be the most appropriate way to collect data from individual sources. However, it is clearly not feasible with a large-scale automated crawl of multiple websites.
2. Another approach is to use only those sources of text that explicitly allow redistribution of contents. Many websites indicate either one of the standard licenses or a custom license in their copyright section. Conveniently, many websites choose to publish their contents under permissive licenses, such as those from the Creative Commons family (see website), or even release the data into the public domain.
3. The data could also be collected without asking for a permission or checking the license. However, a corpus collected in such a manner could not be redistributed for further use by other researchers or the general public. Technically this is achieved by creating a web-interface which allows search through the corpus, but displays only very limited amount of text in the results. The results usually are returned in form of concordance lines, where the search keywords are surrounded by a certain number of words or symbols on both sides.
4. The last option is to redistribute only the links to the original documents, not the text which was extracted. This approach solves the problem of redistribution. However, it raises several issues for the final user of the corpus. First of all, the user would face the inconvenience of having to download and process the entire corpus on his/her own computer, which may be impractical with a reasonably large corpus. Secondly, the web pages may be removed from the web or changed after the original list of links was compiled. This would result in a situation where every user could end up with a slightly different version of the same corpus, raising problems of replicability of the findings.

Many of the serious ethical issues do not apply for gathering of publicly available information on the Internet, as opposed for example to collecting it from interviews with individual respondents, or personal communications. The data have already been made public when they were placed on the web. However general rules of masking personal information and anonymization should still be observed when the data are collected from Internet forums, e-mail groups, or social media websites.

There are also ethical considerations regarding the process of downloading large quantities of data from a single website. For example, an aggressive crawling could slow down the server, thus making it less responsive or even unusable for other users. Larger than usual internet traffic may also cause additional costs for the owner of the website. The best approach would be to consult the owners of the website in question and discuss the most suitable way of collecting the data.

1.3. Use of corpora in Esperanto studies

Esperanto studies is no exception to the general trend in linguistics towards ever wider use of corpora. Dietz (1986), van Oostendorp (1999), and Koutny (2001) were among the first to issue urgent calls for the development and use of Esperanto corpora. Gledhill (2000) responded with a monograph describing Esperanto grammar, based on a corpus of written Esperanto texts comprising 1.5 million words. Wennergren followed suit in 2003, through the support of Esperantic Studies Foundation, with the publically available Corpus of Esperanto, *Tekstaro*, described below in Subsection 2.1. The largest corpus known to this author ever used in research on Esperanto was collected by Eckhard Bick (2007) at the University of Southern Denmark. It consists of 18.5 million words and was used to create and test a Constraint-Grammar-based parser of Esperanto. An expanded version of this corpus is publicly available as the VISL corpus described below in Subsection 2.2. See Table 1 on page 114 for examples of research papers where Esperanto corpora have been used.

2. Readily available Esperanto corpora

2.1. The *Tekstaro de Esperanto*

The easiest way to gain access to a readily available corpus of written original and translated texts in Esperanto is the website of the project *Tekstaro de Esperanto* (*Corpus of Esperanto*; see website). The project was initiated in 2002 by Esperantic Studies Foundation and was implemented by Bertilo Wennergren in 2003. The project is actively maintained to this day.

Contents of the corpus are spread over the entire time span of the existence of the language. The early sources are mostly texts written by the author of Esperanto, L. L. Zamenhof. For example, the earliest item, *La batalo de l' vivo* (*The Battle of Life*) was translated by Zamenhof in 1891—only four years after *La unua libro* (*The First Book*), whose appearance is usually considered to be the beginning of the Esperanto language. Interestingly, *The First Book* is not included into the corpus, as it was not written in Esperanto.

Table 1. Examples of corpora used in studies on Esperanto

Paper	*Description*	*Corpus used*	*Size of corpus*[3]
Tiŝljar (1981)	Book	Spoken Esperanto	24,000
Dasgupta (1993)	Peer-reviewed article	Scientific texts	Unspecified
Schubert (1997)	Book chapter	Distributed Language Translation (DLT) Corpus	Unspecified
Hana (1998)	Master thesis	29 texts consisting for the most part of translated works of literature	460,000
Gledhill (2000)	Monograph	156 texts: journalistic, legalistic, administrative, literary works, personal internet websites	1,563,500
Haszpra (2002)	Book chapter	Esperanto texts	500,000 characters
Minnaja (2002)	Peer-reviewed article	Esperanto texts	Unspecified
Liu (2004)	Peer-reviewed article	*Ekzercaro* de Zamenhof	Unspecified
Herring (2005)	Conference proceedings	Three diachronic samples of journal articles	108,000
Lörnemark (2006)	Grant Proposal for doctoral research	EPAK Corpus (Esperanta PArolingva Korpuso)	100,000
Bick (2007)	Conference Proceedings	Journalistic texts, literature, websites, personal e-mails	18,500,000
Lörnemark (2007)	Grant Proposal for doctoral research	EPAK Corpus (Esperanta PArolingva Korpuso)	100,000
Bick (2009)	Conferene Proceedings	News magazine and literature texts in equal parts	75,000
Krägeloh (2009)	Peer-reviewed article	Articles from *Kontakto* magazine, selected according to specific criteria	3,000 sentences
Dankova (2009)	Peer-reviewed article	Spoken narratives recorded by the researcher	Unspecified

[3]Corpus size is given as number of words, unless other units are specified.

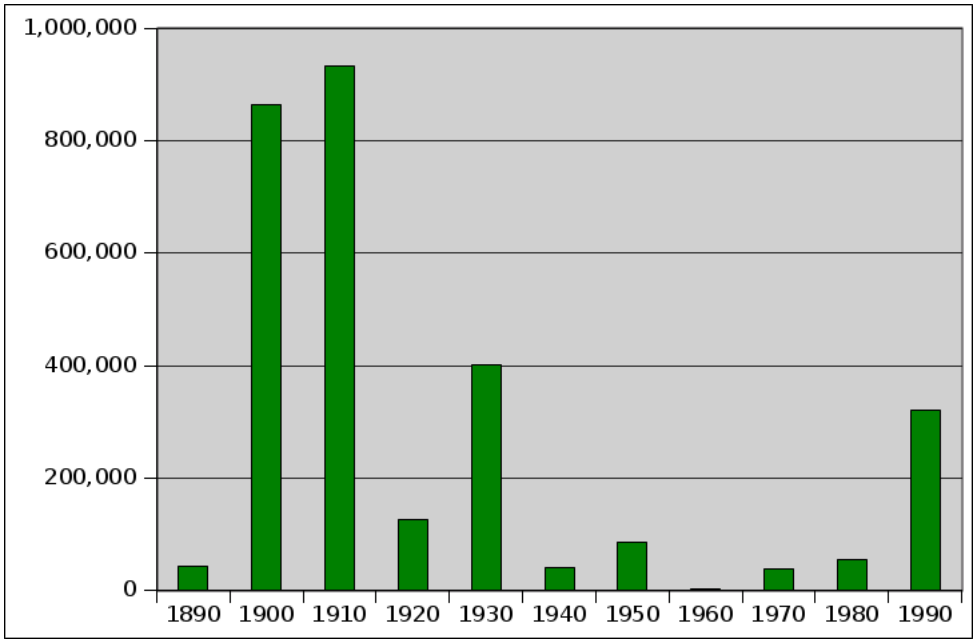

Figure 1. Number of words in *Tekstaro de Esperanto* by decades 1890–1990.

The corpus also includes other core texts of the first years such as *Fundamento de Esperanto* (*Foundations of Esperanto*) and *Fundamenta Krestomatio de la lingvo Esperanto* (*Basic Anthology*).

Presence of the time dimension suggests that *Tekstaro* was probably intended as a monitor corpus. However, due to the relative scarcity of texts from the decades ranging from the 1940s to the 1980s (see Figure 1 above), the corpus could be better described as a diachronic corpus consisting of two major snapshots: (1) early use of Esperanto; and (2) modern use of the language. See the article by Tognini Bonelli (2010:22) for the distinction between monitor and diachronic corpora.

The entire contents of the *Tekstaro* corpus (4,675,412 words—all statistics in this section have been collected in October 2012) is available for search within the website. Most of the texts (totaling 4,524,487 words) of the corpus are also available unabridged, the remaining two books (150,925 words) are available only for search within the website, while complete texts of those books are not available for copyright reasons.

The texts are available in human readable HTML formatting, as well as in XML markup which follows the TEI guidelines (see TEI website). The main use for the markup is to distinguish parts written in Esperanto from those written in other languages, as well as from unesperantized proper names. Markup is also used to retain some formatting of the original texts, for ex-

ample parts highlighted in boldface or italic. Apart from this markup, the corpus is not linguistically annotated, but there are plans to introduce some annotation in the future. However, current corpus search interface allows the use of regular expressions of the Perl programming language and even some language-aware search keywords.

2.2. The VISL Corpus

A large written corpus of Esperanto texts has been collected in the *Visual Interactive Syntax Learning* (VISL) project lead by Eckhard Bick at the Institute of Language and Communication, University of Southern Denmark. The following sources have been used: the news magazine *Monato*, the electronic newsletter *Eventoj*, the Esperanto version of *Wikipedia*, the collection of literature *Elibrejo*, a collection of texts by L. L. Zamenhof, and internet crawls of random websites in Esperanto.

As of October 2012, the VISL corpus has about 58.4 million words. Most of the data are publicly available for search via the Corpuseye web interface (see the Corpuseye page at the VISL website). Permission is needed to access only a small portion of the corpus which contains personal e-mail data (about 120,000 words). The search interface uses Corpus Query Processor (CQP) regular expressions.

The VISL corpus provides part-of-speech, morphological and syntactic annotations generated by an algorithm based on Constraint Grammar.

2.3. Limitations of the existing corpora

One major limitation of the existing corpora is the absence of samples of spoken language either in form of recorded audio or video material, or as transcribed text. Plans to add some spoken material are mentioned in the description of the *Tekstaro* corpus (see the *Priskribo de la projekto* page in the *Tekstaro de Esperanto* website), but at the time of writing no spoken corpus is available. This limitation alone rules out many areas of research.

VISL is the only of the two existing corpora to offer linguistic annotation. The annotations generated by an algorithm based on Constraint Grammar are accurate enough for many practical applications (Bick 2009).

Representativeness is another issue with the available corpora. Neither the *Tekstaro* nor VISL has been designed to represent a balanced sample of language use in the real world. For example, the Lancaster Oslo–Bergen (LOB) corpus, a popular corpus of British English, contains samples from texts of various registers and groups them into 15 categories (see Table 2 on page 117). While the composition of the LOB corpus may not be a blueprint for a perfectly balanced corpus, it still offers some general guidelines. In this

regard, some categories are very poorly, if at all, represented in the existing corpora for Esperanto, for example categories E, H, J, M, N, and R.

Table 2. Text categories and number of samples in the English LOB corpus (Johansson et al. 1986).

	Category in the English LOB corpus	*Number of samples*
A	Press: reportage	44
B	Press: editorial	27
C	Press: reviews	17
D	Religion	17
E	Skills, trades and hobbies	38
F	Popular lore	44
G	Belles lettres, biography, essays	77
H	Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ)	30
J	Learned and scientific writings	80
K	General fiction	29
L	Mystery and detective fiction	24
M	Science fiction	6
N	Adventure and western fiction	29
P	Romance and love story	29
R	Humour	9
	Total	*500*

In the following section I will discuss the publicly available sources of text in Esperanto. Some of the sources are already included in the existing corpora while others could be included to increase the size or to improve the balance.

3. Sources of electronically available text from which to develop written and spoken Esperanto corpora

3.1. Written texts

3.1.1. Works of literature

Project Gutenberg aims to collect and distribute free electronic books. As of October 2012, about 40,000 books have been collected, 84 of them being in Esperanto (see Project Gutenberg website).

All of the books are available in several formats: HTML, EPUB, Plucker, QiOO Mobile, and as plain UTF-8 encoded text, the former being the most desirable for corpus creation. The administrators of Project Gutenberg discourage use of automatic tools to download multiple texts; therefore users are advised to use third party mirror websites for downloads.

The permissive nature of the license (see Project Gutenberg License website) under which the electronic books are distributed, and the fact that most of the books are in the public domain in the United States, leave very little restrictions on their use.

3.1.2. Encyclopediae

A sizable body of Esperanto texts is available in *Vikipedio*—the Esperanto version of *Wikipedia* (see the *Vikipedio* website). As of October 2012, Vikipedio has 169,190 articles in Esperanto with an estimated total of 38 million words (see Wikipedia Statistics Esperanto website).

The entire database of Wikipedia in Esperanto is available for downloading (see Wikimedia Downloads website). Third party tools may be used to extract contents of the database and render it to plain text (see Wikipedia: Database download website). It should be noted that users are discouraged from using web crawling tools to download large numbers of articles directly from the website.

Great care must be taken to filter out machine-generated stub articles, templates and other components of the website, as their presence in a corpus may abnormally raise occurrences of some words or phrases.

Articles in Vikipedio are distributed under Creative Commons Attribution-ShareAlike 3.0 Unported license (see the Creative Commons website).

3.1.3. Scientific and technical texts

A small body of scientific texts in Esperanto can be obtained from 5 issues of *Esperantologio / Esperanto Studies* journal (see website). As of October

2012, 464 pages of scientific papers have been published.[4] While most of the articles are in Esperanto, there are also some in English, French, and German (see page 44). Issues of *Esperantologio / Esperanto Studies* are available in electronic form: issues 1–2 (1999–2001) in DVI, PS, and PDF formats; issues 3–6 (2005–2013) in PDF format.

Scienca kaj Teknika Esperanto-Biblioteko (*Scientific and Technical Esperanto Library*) aims to collect scientific and technical documents written in Esperanto (see the STEB website). The documents have been collected from very diverse sources and from various fields, which makes it an especially valuable source of text samples in this register.

3.1.4. Official documents

The *Akademio de Esperanto* (Academy of Esperanto) aims to ensure that change of Esperanto language stays within the spirit of the foundation of the language. Although decisions of the Academy have only advisory status, they are often regarded as the official position on linguistic questions regarding Esperanto. Over the years, the Academy of Esperanto has generated a sufficiently large body of documents, most of which are publicly available online (see the *Akademio de Esperanto* website). As the documents are written in a rather official and scholarly manner, they could be a good candidate for samples of official documents. The disadvantage is that the documents discuss quite a narrow range of topics—mainly Esperanto language and to some extent organizational affairs of the academy.

The *Esperanta Civito* (Esperanto Citizens' Community) aims to create a community of Esperanto speakers, recognized by international law. Goals and ideas behind the organization are controversial among other Esperantists and have often been criticized (see *La liturgio de l' foiro* website) and even mocked (see the *Esperanta Respubliko* website). However, members of the organization have created a significant body of official documents in Esperanto, which are probably as close to official government documents as one could get in this language. For example, the organization has a constitution, laws, and various declarations. It also has an organizational structure which resembles a state: a consul, parliament, court and committees of various sorts. It is hard to judge how active these organizations in fact are. However documents created by them may be a unique contribution to the variety of contexts in which Esperanto is used—in this way forming a valuable addition to a balanced corpus of Esperanto texts. Some of the documents are available on the website of Esperanto Citizens' Community in HTML format (see the *Esperanta Civito* website).

[4] With the present issue, the number of pages is 596.

Due to longstanding relations between the Esperanto movement and the United Nations, many important UN-related documents have been translated into Esperanto (see the *Esperanto por UN* website).

Esperanto organizations should also have many documents not intended for the public: foundation documents, reports and internal or external communications of various sorts. Provided that the documents do not include too sensitive information, they might be made available when requested.

3.1.5. Periodicals, news websites and blogs

Monato is a monthly news magazine in Esperanto. It differs from other periodicals as it publishes articles in Esperanto language but not about the movement or linguistic aspects of the language. Because of this restriction, Monato is a very welcome source of texts, as it covers a much larger range of topics by a wider range of authors. Articles from 2003 till 2011 are freely available in HTML formatting (see the *Monato* website). Articles from 1997 to 2003 are already included into the *Tekstaro* corpus (totaling 578,826 words). The entire archive of the articles currently available in electronic form from 1997 to 2011 should exceed the mark of 1 million words. The size and diversity of content combined with high editorial standards make Monato one of the most desirable sources for construction of corpora. For this reason articles from Monato are already present in both the *Tekstaro* and VISL corpora.

La Ondo de Esperanto is another monthly periodical in Esperanto which has an extensive database of articles that have appeared in the magazine since 1999. The topics mostly cover the Esperanto movement in Russia and abroad. The magazine is also noteworthy because it publishes original and translated literature in Esperanto (see *La Ondo de Esperanto* website).

Libera Folio is an independent bulletin about the Esperanto movement. As of October 2012, 1,741 articles of varying length have been published. Research by this author allows to estimate the total size of all articles to about 1.3 million words. The website offers a complete list of articles in a well organized archive section (see the *Libera Folio* website).

Other sources of larger collections of texts available in Esperanto are:

- Archives of *Revuo Esperanto*, the official publication of the Universal Esperanto Association, offer collections of published articles from 2002 till 2005 (see UEA website).
- *Vikinovaĵoj*, the Esperanto version of Wikinews has archives from 2010 till 2012 (see the *Vikinovaĵoj* website).
- The Esperanto version of *Le Monde diplomatique* offers a complete and freely accessible archive of articles (starting from 1999 and totaling 1.6

million words as of March 2013) on its website (see *Le Monde diplomatique* website).

- Articles of the discontinued newsletter *Eventoj* from 1999 till 2002 are available for downloading (see the *Eventoj* website)

There are numerous blogs written by Esperantists. A few of them are notable because of the size of their archives:

- Blog of Esperanto-USA (see Esperanto-USA website)
- *Neniam milito inter ni* (see website)
- *Eŭropa Civitano* (see website)
- *Bagateloj* (see website)

3.1.6. Internet forums, social media websites

One of the most active internet forums in Esperanto is the forum part of *Lernu!* (Learn!) website (see Forum page on the *Lernu!* website). Because the website is intended for students of the language, discussions in its forums could be used to create a learner corpus. As of October 2012, there are 173,347 messages in *Lernu!* forums (see Information about the forums page on the *Lernu!* website).

Esperanto is widely used in most popular social media networks. There is a significant community of Esperanto speakers on *Facebook* and *Twitter*. *Ipernity* is especially known for its popularity among Esperantists. *Esperanto.com* is a social media website dedicated specifically to the Esperanto community.

3.2. Oral Texts

3.2.1. News broadcasts

The website of *Radio Verda*, a free internet radio in Esperanto, has 197 (as of February 2013) audio clips of spoken Esperanto in its archives (see the *Radio Verda* website). For the most part the clips contain entertaining stories on various topics read by Aaron Chapman and Charlene Daley. Some broadcasts also include spoken recordings sent by the listeners. No transcriptions of the broadcasts are available.

The archives of *Svisa Radio Internacia* are also available on the Internet (see the *Svisa Radio Internacia* website). Transcriptions of some of the broadcasts are also available.

There are other news services in Esperanto, but their archives are not freely available on the Internet. For example, *Radio Televido Esperanto* (see RTE website) and *Radio Vatikana* (see website).

3.2.2. Speeches

Many speeches by speakers including prominent figures of the Esperanto movement can be found on *YouTube*. For example, a speech by Humphrey Tonkin at an event at the North American Summer Esperanto Institute and the opening speech by Probal Dasgupta at a World Congress of Esperanto (see YouTube website for both links).

3.2.3. Films

The Universal Language is a documentary by Sam Green (see The Universal Language website) about Esperanto language and the movement related to it. The film features several interviews with Esperantists. Unfortunately, the parts of the interviews which made it into the final cut of the film may not be long enough to make a significant contribution to a corpus of spoken language. However, complete interviews, especially if there are more of them than appeared in the film, could be of interest for this purpose.

The horror film *Incubus* (see IMDb website) directed by Leslie Stevens is probably the best known film ever made in Esperanto. However, due to poor pronunciation by the actors it hardly offers any value for linguistic research.

The short film *Senmova* (see Tuğçe Sen interview in the *Libera Folio* website) also contains several dialogues in Esperanto.

3.2.4. Conversations

Naturally occurring conversation is probably the most organic use of language possible; therefore it is a very important register to collect corpora from.

Some examples of spoken conversations in Esperanto could be found on video hosting websites (for example, see Part 1, Esperanto Encounter video on YouTube website).

Instant messaging is the form of communication over the Internet closest to spoken conversation. Logs of Internet Relay Chat (IRC) and other instant messaging services could be used to gather samples.

4. Sources of print text from which to develop corpora

Since the creation of the language, a variety of books have been written in or translated into Esperanto. Over the decades, this amounts to a large body of literature. For example, the Hector Hodler Library of the Universal Esperanto Association has about 15,000 books and a collection of periodicals in Esperanto (see Esperantic Studies Foundation website). Another noteworthy example is the International Esperanto museum at the National library of

Austria, because it already offers part of its archives in electronic form (see the *Österreichische Nationalbibliothek* website).

Butler Library at the Esperanto Association of Britain, *Centre de documentation et d'étude sur la langue internationale* at the public library in La Chaux-de-Fonds, Switzerland, and several other libraries also have sizable collections of publications in Esperanto. The total number of existing publications is estimated to be between 25,000 and 40,000; however, these numbers proved to be difficult to verify.

Converting all these books, periodicals, and other documents into digital media would be a tremendous project, out of scope of an individual researcher. However, with steady effort and support from major Esperanto organizations (for example Universal Esperanto Association, Esperantic Studies Foundation or local organizations) this could be done, offering an invaluable collection of texts for Esperanto corpora of the future.

5. Discussion and conclusions

The *Tekstaro* and VISL corpora certainly offer some data to work on for a researcher interested in linguistic aspects of Esperanto. However, the corpora fall behind present-day levels at least in quantitative terms. There are also issues with the balance of the corpora. Both these aspects could be improved by collecting additional samples from the registers discussed in Section 3 above.

Probably the best solution for Esperanto studies would be creating a large and well-balanced reference corpus, which could be versatile enough for many different applications. Even when a smaller specialist corpus is more appropriate for a particular research question, it may be much easier to create a sub-corpus from the existing corpus than to create a new corpus from scratch. A single reference corpus could also facilitate replicability of findings by other researchers.

Absence of samples of spoken language use is another serious shortcoming of the existing corpora. This may be explained by the fact that such corpora are much more difficult to create compared to corpora of written language.

The accuracy of machine-generated annotations of the VISL corpus may be high enough for many practical applications. However, a small manually annotated corpus could still be useful. For example, it could be used as a gold-standard benchmark for testing of linguistic analysis algorithms. In turn, the tools trained on the manually annotated corpus could be used in the future to annotate the rest of the corpora, or to provide the same functionality on-the-fly.

Lernu! forums are a particularly interesting source of text which could be harvested to create a learner corpus with a relatively low amount of effort.

Accompanied by users' self-reported level of language proficiency and if possible also by test results, such a corpus could be useful even beyond studies of Esperanto: It could as well help to reveal valuable insights into the process of L2 language acquisition.

The period of relatively scarce amounts of text during the decades from the 1940s to 1980s in the *Tekstaro* corpus raises certain questions. Namely, does this indicate a period of relatively low activity in the Esperanto movement, or is it merely an artifact of the way the corpus was sampled? A bibliographic study of quantities of literature published throughout the history of Esperanto could shed some light on this question.

Interestingly, there is only one instance (at least to this author's knowledge) when existing corpora were used for actual research on Esperanto (Bick 2009), while there were several studies where authors chose to compile corpora themselves (see Table 1 on page 114). Of course there are good reasons to use a custom specialist corpus in some cases, but the general tendency seems to suggest that there may be some obstacles related to using the existing corpora. For example, other researchers may not be aware they exist or know where to find them, or perhaps their search interface may not be optimal. By raising awareness of existing corpora through published research on them, as this article has attempted to do, and by working to upgrade search interfaces, such obstacles can easily be overcome.

Acknowledgment

I am grateful to three anonymous referees, whose helpful remarks have enabled me to substantially improve the text.

References: Books and articles

Bick, Eckhard. 2007. Tagging and Parsing an Artificial Language: An Annotated Web-Corpus of Esperanto. **In:** *Proceedings of the Corpus Linguistics Conference CL2007*, University of Birmingham, UK 27–30 July 2007. Edited by Matthew Davies, Paul Rayson, Susan Hunston, Pernilla Danielsson. Available at `http://ucrel.lancs.ac.uk/publications/CL2007/` (accessed 2013-03-07).

Bick, Eckhard. 2009. A Dependency Constraint Grammar for Esperanto. **In:** *Constraint Grammar Workshop at NODALIDA 2009, Odense*. NEALT Proceedings Series, Volume 8, pp. 8–12. Tartu: Tartu University Library. Available at `http://beta.visl.sdu.dk/pdf/cg-workshop2009_dep.pdf` (accessed 2013-03-07).

Dankova, Natalia. 2009. Temporality in spoken Esperanto. *Esperantologio / Esperanto Studies* **4**, 43–60.

Dasgupta, Probal. 1993. Idiomaticity and Esperanto Texts: An Empirical Study. *Linguistics* **31**.2, 367–386.

Dietz, Joachim. 1986. Projekt der rechnergestutzten Erarbeitung eines Worterbuches von Esperantowurzeln [A Project of a Computer-Assisted Compilation of a Dictionary of Esperanto Roots]. *Wissenschaftliche Zeitschrift der Martin Luther Universität Halle-Wittenberg, Gesellschafts-Sprachwissenschaftliche Reihe* **35**.5, 90–91.

Fiedler, Sabine; Liu, Haitao (Eds.). 2001. *Studoj pri Interlingvistiko/Studien zur Interlinguistik; Festlibro Omaĝe al la 60-Jariĝo de Detlev Blanke/Festschrift für Detlev Blanke zum 60. Geburtstag*. Prague: KAVA-PECH.

Gledhill, Christopher. 2000. *The Grammar of Esperanto. A Corpus-Based Description*. München: Lincom Europa. PDF edition. (Reviewed by Haitao Liu in *Esperantologio / Esperanto Studies* **2** (2001), 39–47.)

Hana, Jiří. 1998. *Two-level morphology of Esperanto*. Prague: Master Thesis at MFF UK Praha. `www.ling.ohio-state.edu/~hana/esr/thesis.html` (accessed 2013-03-07).

Haszpra, Ottó. 2001. Liter-ofteco en Esperantaj tekstoj. [Frequency of Letters in Esperanto Texts]. **In:** Fiedler & Liu (Eds., 2001:346–364).

Herring, Joshua. 2005. *Syntactic and Lexical Changes in Esperanto: A Quantitative and Corpus-Based Survey*. `http://cllt.osu.edu/mclc/paper/syntactic_herrring.pdf` (accessed 2013-03-07).

Johansson, S.; Atwell, E.; Garside R.; Leech G. 1986. The tagged LOB corpus: Users' Manual. `http://khnt.hit.uib.no/icame/manuals/lobman/INDEX.HTM` (accessed 2012-10-25).

Koutny, Ilona. 2001. Defioj de moderna leksikografio por Esperanto [Challenges of Modern Lexicography to Esperanto]. **In:** Fiedler & Liu (Eds., 2001:660–674).

Krägeloh, Christian U. 2009. Lingva fono kaj uzado de frazstrukturoj en skribita esperanto. *Esperantologio / Esperanto Studies* **4**, 31–42.

Liu, Haitao. 2004. La lingvistikaj konceptoj de Zamenhof [Linguistic Concepts of Zamenhof]. *Grundlagenstudien aus Kybernetik und Geisteswissenschaft (GrKg/Humankybernetik)* **45**.4 (Dec. 2004), 155–165.

Lörnemark, Christer (2006). An Esperanto spoken language corpus: A proposal for a pilot study. Grant proposal submitted to Esperantic Studies Foundation.

Lörnemark, Christer (2007). An Esperanto spoken language corpus: A proposal for further study. Grant proposal submitted to Esperantic Studies Foundation.

McCarthy, Michael; O'Keeffe, Anne. 2010. Historical perspective: what are corpora and how have they evolved? **In:** O'Keeffe & McCarthy (Eds. 2010).

McEnery, Tony; Hardie, Andrew, Eds. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press. Kindle edition.

Minnaja, Carlo. 2002. Anaphora with Relative Pronouns: An Algorithm for Italian and Esperanto. Part II: The Test for Esperanto. *Grundlagenstudien aus Kybernetik und Geisteswissenschaft (GrKG/Humankybernetik)* **43**.3, 115–124.

Nelson, Mike. 2010. Building a Written Corpus: What are the Basics? **In:** O'Keeffe & McCarthy (Eds. 2010).

O'Keeffe, Anne; McCarthy, Michael. Eds. 2010. *The Routledge Handbook of Corpus Linguistics*. London: Routledge. Kindle edition.

van Oostendorp, Marc. 1999. Komputado ŝanĝas ĉion [The Computer Changes Everything]. *Esperanto* **92**.4 (April 1999), 68.

Schubert, Klaus. 1997. Ausdruckskraft und Regelmassigkeit: Was Esperanto für automatische Übersetzung geeignet macht [Expressiveness and Regularity: Why Esperanto is Suited for Automatic Translation]. **In:** Tonkin, Humphrey (Ed.), *Esperanto, Interlinguistics, and Planned Language*, pp. 117–139. Lanham, MD: University Press America.

Tišljar, Zlatko. 1981. *Frekvencmorfemaro de parolata Esperanto*. Zagreb: Internacia Kultura Servo.

Tognini Bonelli, Elena. 2010. Theoretical Overview of the Evolution of Corpus Linguistics. **In:** O'Keeffe & McCarthy (Eds. 2010).

References: Websites

Akademio de Esperanto. Aktoj de la Akademio. `www.akademio-de-esperanto.org/aktoj/index.html` (accessed 2013-02-10).

Bagateloj. `http://bagateloj.org/archive` (accessed 2013-02-12).

Creative Commons. Attribution-ShareAlike 3.0 Unported. `http://creativecommons.org/licenses/by-sa/3.0/legalcode` (accessed 2013-01-28).

Esperanta Civito. Normaro. `www.esperantio.net/index.php?id=14` (accessed 2012-10-25).

Esperanta Respubliko. `www.esperantarespubliko.blogspot.com/` (accessed 2012-10-25).

Esperantic Studies Foundation. Research Tools. `http://esperantic.org/en/research/tools` (accessed 2013-02-13).

Esperanto por UN. Tradukitaj UN-dokumentoj. `https://sites.google.com/site/esperantoporun/dokumentoj/traduko-de-un-dokumentoj` (accessed 2013-02-24).

Esperantologio / Esperanto Studies. `www2.math.uu.se/esperanto` (accessed 2013-01-28).

Esperanto-USA. Blogs. `http://esperanto-usa.org/blog/` (accessed 2013-02-12).

Eŭropa Civitano. `http://europacivitano.wordpress.com/` (accessed 2013-02-12).

Eventoj. Arkivo de la gazeto Eventoj. `www.eventoj.hu/arkivo/` (accessed 2013-02-12).

IMDb. Incubus. `www.imdb.com/title/tt0059311/` (accessed 2013-02-16).

La Ondo de Esperanto. Gorecka, Halina & Korĵenkov, Aleksander, Eds. Iom pri nia revuo. `http://esperanto.org/Ondo/Pri_lode.htm` (accessed 2013-02-12).

Le Monde diplomatique. Arkivoj. `http://eo.mondediplo.com/archives.php3` (accessed 2013-02-12).

Lernu!. Forum. `http://en.lernu.net/komunikado/forumo/index.php` (accessed 2013-02-11).

Lernu!. Information about the forums. `http://en.lernu.net/prilernu/statistiko/lernuforumoj2.php` (accessed 2013-02-11).

Libera Folio. `www.liberafolio.org/` (accessed 2013-02-12).

Libera Folio. Tuğçe Sen: En mia koro kaj menso mi naskiĝis esperantisto. `http://liberafolio.org/2010/senmova` (accessed 2013-02-12).

La liturgio de l' foiro. Camacho, J. 1999. `http://esperanto.net/literaturo/tekstoj/camacho/liturgifoir.html` (accessed 2013-02-25).

Monato. Indeksoj de Monato. www.esperanto.be/fel/mon/moninde.php (accessed 2013-02-25).

Neniam milito inter ni. http://neniammilitointerni.over-blog.com/ (accessed 2013-02-12).

Österreichische Nationalbibliothek. Gescannte Zeitschriften online. www.onb.ac.at/sammlungen/plansprachen/19056.htm (accessed 2013-02-24).

Project Gutenberg. www.gutenberg.org/browse/languages/eo (accessed 2013-01-28).

Project Gutenberg. Project Gutenberg License. www.gutenberg.org/wiki/Gutenberg:TheProjectGutenbergLicense (accessed 2013-01-28).

Radio Vatikana. www.radiovaticana.va/esp/index.asp (accessed 2013-02-25).

Radio Verda. http://radioverda.com/rv100plus/ (accessed 2013-02-14).

RTE. www.livestream.com/rteradiotelevidoesperanto (accessed 2013-02-25).

STEB. www.eventoj.hu/steb/ (accessed 2013-02-25).

Svisa Radio Internacia. http://esperanto-gacond.ch/radioprelegoj.html (accessed 2013-02-25).

TEI: Text Encoding Initiative. www.tei-c.org/index.xml (accessed 2013-02-09).

Tekstaro de Esperanto. http://tekstaro.com/ (accessed 2013-02-09).

Tekstaro de Esperanto. Priskribo de la projekto. www.tekstaro.com/tekstaro.html#priskribo (accessed 2013-02-24).

The Universal Language. http://esperantodocumentary.com/en/about-the-film (accessed 2013-02-16).

UEA. Revuo Esperanto. www.uea.org/revuo/index.html (accessed 2013-02-12).

Vikinovaĵoj. http://eo.wikinews.org/ (accessed 2013-02-12).

Vikipedio. http://eo.wikipedia.org (accessed 2013-01-28).

VISL research and development project at the Institute of Language and Communication, University of Southern Danmark. Corpuseye. http://corp.hum.sdu.dk/cqp.eo.html (accessed 2013-02-07).

Wikimedia. Wikimedia Downloads. http://dumps.wikimedia.org/eowiki/ (accessed 2013-01-28).

Wikimedia. Wikipedia Statistics Esperanto. http://stats.wikimedia.org/EN/TablesWikipediaEO.htm (accessed 2013-01-28).

Wikipedia. Wikipedia: Database download. http://en.wikipedia.org/wiki/Wikipedia:Databasedownload (accessed 2013-01-28).

YouTube. Part 1. Esperanto Encounter. www.youtube.com/watch?v=5swk95Fg1fM (accessed 2013-02-25).

YouTube. Parto 3a. Humphrey Tonkin: Esperanto: Pasinteco, Nuntempo, Estonteco. www.youtube.com/watch?v=P2naAiifXzA (accessed 2013-02-25).

YouTube. Universala Kongreso de Esperanto 2011, #04. www.youtube.com/watch?v=6KI9bAN2b8 (accessed 2013-02-25).

Vilius Normantas is a software engineer who specializes in solutions relating to active trading in the financial markets. He holds a Bachelor degree in Computer Science and at present is a PhD student at the Institute of Mathematics, Tajik Academy of Sciences. Vilius hopes to merge his computer programming skills with his general interest in languages to make a meaningful contribution in the field of Information Theory.

Author's address: Institute of Mathematics, Dushanbe, Republic of Tajikistan, *or*
Kranto 17-oji g. 17, LT-45292 Kaunas, Lithuania
E-mail: vilius@norma.lt

Received 2012-10-29.
Revised version received 2013-02-26.
Accepted for publication 2013-03-08