THE TEXT HAS BEEN TRANSLATED FULLY AUTOMATICALLY

Automatic Documentation and Mathematical Linguistics *Nauchno-Tekhnicheskaya Informatsiya, Seriya 2*
Vol. 37, No. 3, pp. 35-42, 2003 *No. 6, pp. 37-43, 2003*
UDC 81:001.103](480)

CORPUS LINGUISTICS IN FINLAND: A RESOURCE SURVEY*

M. V. Kopotev (Helsinki)

An indication is given of the main computer linguistic resources set up in Finland and the ways in which they may be used. The survey is descriptive and is certainly only fragmentary. Over 40 different-language corpuses are briefly characterized, with references to the corresponding publications where one can find more detailed information.

Finnish corpus linguistics and computer linguistics generally has an ancient tradition, which gives it authority in the world community and has produced solid results in various areas. The first projects for electronic corpuses appeared in the 1960s, as in many other countries [1, 2]. From the start, this line in Finland was closely related to the writing of original computer programs for processing text, as well as close international links and devotion to current topics in lexicography and grammatical description ([3-6]; see also the round-table material on corpus linguistics in "Korpuslingvistiikan työpaja l: Korpukset ja ohjelmat", pp. 126-134 of [7]). The major feature of computer linguistics in Finland has become the close connection with the writing of end-user products, which has included collaboration with commercial firms [8; 1, pp. 50-54 and 62-64].

This paper is of information type and has particular purposes such as giving Russian linguists a conception of the main computer linguistic resources in Finland and determining the scope for them to use them. Each existing corpus is indicated as regards position at the present time, which is reduced in some cases to indicating the place of creation and initial storage. The characteristics of each are indicated by listing the places of detailed description (on the Internet and/or as a paper publication), in which full information can be obtained.

Many of the resources described below provide remote access to the files (most of the servers work under the control of the Unix OS, which in general involves the user's machine having Unix-Client, e.g., the program F-Secure SSH-Client). I do not discuss in detail the technical and organizational aspects of access and merely state that almost all of them are accessible for free use for research and teaching purposes. In most cases, this requires one to obtain permission from the administrator or owner of the corpus. Contact information is given on the corresponding Internet sites or in articles on the topic.

The following comment is important. We are concerned with a definition of the corpus content. There are multiple meanings or uncertain use of this term, which lead to some general tendency for the name electronic *corpus* to be given to any collection of texts put into digital format. On the other hand, recently the term corpus has increasingly been used not simply for text (English running text) but linguistic material especially selected on ceratin principles.

"So a corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration" [9, p. 24].

However, in spite of the expansion of the new approach, old corpuses (i.e., simply electronic texts) still retain their linguistic value in many areas. This is dependent on the substantial differences in quantity and quality of the work done. For example, the last number of text collections in English poses substantially more complicated tasks (various types of annotation, parallel corpuses, speech records presented in electronic form, and so on). On the other hand, in many modern languages there are as yet no simple well-balanced representative corpuses, quite apart from annotated ones. Special and equally difficult problems arise for the creation of any corpus of ancient texts. The

*The author is indebted for comments made on the draft by Prof. F. Karlsson, Prof. A. Mustaioki, and Dr. A. Nikunlassi.

significance of a corpus is in fact determined not only by the extent to which the material has been processed, but also by its specific weight, which is dependent on the representativeness of the material in the corpus sector of world linguistics. Therefore, the present survey contains not only any text corpus with a branched and detailed system of annotation but also sometimes just a simple collection of texts having linguistic value by virtue of the uniqueness of the material.

FINNISH IT CENTER FOR SCIENCE (CSC)
http://www.csc.fi

CSC is a firm belonging to the Finnish Education Ministry, which provides services on storing and processing data for various commercial and scientific bodies, and also for individual researchers. This center is equipped with powerful supercomputers and occupies a special position among such firms in Finland and the world generally*. The center provides researchers with managed specialties (physics, chemistry, genetic engineering, and so on) and scope for working with databases and software belonging to CSC and other firms not only on the center's computers but also by network access. As regards linguistic tasks, there have been over 60 projects at the CSC, most of them concerned with corpus linguistics.

The CSC server contains one of the largest collections of electronic texts in Finland: the Kielipankki or Language Bank (http://www.esc.fi/kielipankki), which contains various corpuses, word indexes, speech recordings, and so on. The total volume of the bank is 747,236 documents, total number of words 210,539,150. About 86% of documents are in Finnish (total volume 179,602,389 words).

Brief information is given below on the main collections (general information on the corpuses and programs accessible at the CSC center can be found in [10]).

Suomen kielen tekstipankki (SKTP)
(Finnish Language Text Bank)
http://www.csc.fi/sktp

At the start of 2001, the SKTP text bank contained Finnish texts of total volume over 180 million word uses and the texts of Finnish Swedes (about 35 million word uses in 2000). The texts are annotated in the SGML format in accordance with the TEI standards and consist of several parts: newspapers, journals, novels, and scientific articles (1990-2000). About half of the texts are supplied with syntactic and morphological information inserted by the TEXTMORFO and SWESG programs. The composition of the corpus can be found at the address http://www.csc.fi/skpt/sktp-tekstit .html. Information is given on the type of annotation at the address http: //www.esc. fi/kielipankki/aineistot/naytaAineistot .phtml.

About half of the Finnish texts and a third of the Swedish ones constitute the SKTP-LT lexical subcorpus ("Suomen kielen tekstipankin leksikaalinen tietokanta"). This is an index database providing various forms of information on wordforms or lexemes (morphosyntactic features, frequency, initial form for wordforms, total frequency for all wordforms of a given lexeme, number of wordforms for a lexeme, number of files, which is approximately equal to the number of storage units, books or articles, in which the units are encountered). The subcorpus is served by the LEMMIE program (see [11] for more details).

Oulun korpus
(Oulu Corpus [12])
http://www.csc.fi/kielipankki/iulkaisut/opas/c623.phtml#AEN625

The corpus was prepared at the start of the 1980s at the Department of Finnish and Saamian in the University in Oulu and includes 429,058 words from 5800 texts collected from literary products, recordings of radio transmissions, advertisements, newspapers, and journal articles in Finnish. In 1997, the corpus was transferred to the SGML format and is now accessible on the CSC server. The corpus has been grammatically annotated by means of the CQP (Corpus Query Processor).

*An example is provided by the characteristics of the Cedar computer, which linguists can use. Number of processors 128, theoretical peak processing rate 76.8 gigaflops a second, RAM 160 Gbyte. (http://www.csc.fi/metacomputer/laitetiedot .html .en)

Suomen murteiden morfologinen digitaaliarkisto (Finnish dialect electronic morphological archive)
http://www.ioensuu.fi/fld/methodsxi/abstracts/kukkola.html

A morphologically annotated archive of Finnish dialects has existed since 1967 and includes about 2 million units illustrating almost 1000 features of the Finnish dialect morphological system. In recent years it has been gradually converted to electronic form in the SGML format. It has been prepared at the Department of Finnish Language at Helsinki University.

Keskiranskan korpus
(Middle French language corpus)
http://www.csc.fi/kielipankki/iulkaisut/opas/x693.phtml

This was set up in the Department of Romance and Classical Languages at Juvaaskul University and includes Middle French texts from the 14th to 16th centuries. The total volume of the 29 texts is about 1 million words. Out of them, 14 texts are accessible on the CSC server (430 thousand words). The corpus contains various stylistic products and prosaic texts: novels, songs, and chronicles. The corpus can be used for lexicological and syntactic purposes.

LE-PAROLE
http://www.hltcentral.org/usr docs/proiect-source/parole/ParoleFinal.pdf

An international project supported by the European Union with the purpose of preparing a corpus for 11 European languages on unifying principles. The Finnish part has been set up jointly by the Division of General Linguistics at Helsinki University and the Language Research Center of Finland and contains about 20 thousand words annotated morphologically and syntactically. In addition to the Finnish corpus PAROLE-FI, which has been constructed under the project, the CSC server gives access to the German part of the international project (PAROLE-GE).

In addition to these projects carried out by Finnish experts, the CSC center has various corpuses set up in other countries. I list the main ones with references to the relevant pages.

Susanne corpus (http://www.cogs.susx.ac.uk/users/geoffs/SueDoc.html).
CHILDES (http://childes.psv.cmu.edu/).
The Oxford Text Archive (http://ota.ahds.ac.uk/).
Le Monde (http: //www. icp. inpg. fr/ELRA/).
WordNet (http://www.cogsci.princeton.edu/ ~ wn/).

The University of Helsinki language corpus server (UHLCS)
http://www.ling.helsinki.fi/uhlcs

The Helsinki University language server is a multilanguage databank administered by members of the General Linguistics division. It was set up in 1980 and originally contained materials on Finnish, Swedish, and English. The main server was set up in the creation of one of the first Finnish corpuses, the HKV corpus, which is a syntactically annotated corpus of Finnish (see [13] for more details of this and some analysis results). Since the 1980s, the database has been supplemented with projects performed by various Helsinki University divisions ("Databank for endangered Finno-Ugric languages", "LENCA-group project", "A Finland Swedish Corpus (FISC), etc.), and also by annotating materials prepared in other countries. Large parts of these annotated corpuses are due to application of the theory: Two-Level Grammar [14] and Constraint Grammar [15], together with software based on them for automatic morphological and syntactic analysis http://www.lingsoft.fi/doc.

For example, the corpus for Finnish, Swedish, and Swahili provides examples of what has been done at Helsinki University. There is a series of subcorpuses for Russian texts, which provide a good example of the work done by several research groups. Some of them contain annotation (syntactic, morphological, and lexical-semantic), while some are lemmatized. There is also a special corpus containing lists of words (without context) for several languages (Finnish, Dutch, French, German, Italian, Norwegian, and Swedish). The UHLCS server now includes language materials for more than 50 languages (data on the languages represented on the server can be found from

the address http: //www JI ing .helsinki. fi/uhlcs/data/languages .html.

In 2000, parts of these corpuses were prepared for general use for research and teaching purposes. The Division of General Linguistics resembles many other research centers in providing software and tools for processing texts (see http://www.ling.helsinki.fi/uhcls/tools/tools.html).

The corpuses accessible on the UHLCS server are linked into two large groups: Helsinki Corpora I and Helsinki Corpora II.

Helsinki Corpora I
http://www.ling.helsinki.fi/uhlcs/data/helsinki"Corpora-Lhtml

This part contains corpuses for Finnish (some of the texts with morphological annotation, some with phonetic transcription), English (some of the texts with morphological annotation, some with phonetic transcription), Swedish (some of the texts with morphological annotation), Swahili, Russian, Latin, German, Greek, and other languages of total volume several million words.

Helsinki Corpora II
http://www.ling.helsinki.fi/uhlcs/data/helsinki-corpora-n.html

This collection contains materials prepared by several research groups. It covers languages of the former Soviet Union (Uralian, Caucasian, Indo-European (Iranian, Slavonic (only Ukrainian)), Chukotka-Koryak, Tungus-Manchurian, and Turkic). Particular value attaches to materials prepared under the "Databank for endangered Finno-Ugric languages" (morphologically, syntactically, and lexicosemantically annotated texts from the Finno-Ugric languages of the former Soviet Union [16]).

Special interest for the Russian reader attaches to a corpus of Russian texts. The UHLCS server contains materials of volume more than 5 million word uses prepared by research groups and individual researchers (http://www.ling.helsinki.fi/uhlcs/readme-all/README-Russian.html). All the texts are presented with SGML annotation.

Annotated Russian language corpus: this consists of the texts of journal publications in 1999 and 2000 chosen by the translating division at Tampere University. The corpus contains almost 2 million text forms morphologically annotated by means of the RUSTWOL program [17].

Newspaper article corpus: prepared by the Lingsoft company, an unannotated corpus containing the texts of articles of volume about 200 thousand words.

Corpus of conversational Russian: prepared by the Russian Language Institute, RAS (Moscow). The corpus has been translated into the SGML format in the General Linguistics section of Helsinki University.

Professor J. Fowler's unannotated corpus of Russian literary texts (among others, there are some translations into English). Volume almost 3 million words.

Russian journal texts (journals "Novoe Vremya", "Ogonek", and "Sputnik"). Unannotated corpus of volume about 100 thousand words.

Ryscard database: prepared by Uppsala University. Unannotated texts of volume almost 400 thousand words.

Uppsala text corpus. Copy of a certain unannotated corpus, containing about 1.5 million words.

Finland Languages Research Center
http://www.kotus.fi/inenglish/

The center (Kotimaisten kielten tutkimuskeskus, KOTUS) was founded by the Finnish Ministry of Education as the leading institute for researching the languages of the indigenous peoples of Finland. The sphere of interests covers primarily Finnish, Swedish, Saamian, and Tsygan, and the Finnish language of records. The main spheres of activity are preparing dictionaries, collecting and processing language materials, and participating in language planning. The center possesses very large dictionary card indexes to its various lines of activity. As regards electronic materials, the total volume of those in digital form is over 76 million words in various types of electronic collection. Some of the prepared materials are located on the CSC server (see above), so I consider only what is directly accessible in the center.

Dictionary corpus: this is the basis for the interpreting dictionary of modern Finnish [18] and is a dictionary card index drawn up from newspapers, literary products, radio and TV transmissions, and also Internet publications.

Since 1987, an electronic form has been set up for the card indexes, and about 100 thousand dictionary articles have so far been translated into electronic form (SGML format).

Card index dictionary of modern Finnish: this has been translated into electronic form and contains about 2,100,000 words, including compound words, together with materials for a dictionary of new words and meanings (for 1950-1980).

19th century electronic archive: this includes unannotated texts of about 100 books published between 1810 and 1900 (text list accessible from the address http://www.kotus.fi/aineistot/1800/1800 sahkoisetanaineisto. shtml).

Archive of electronic texts from old Finnish. The corpus contains major written records from the 16th-18th centuries, total volume over 3.2 million words (list accessible on http://www.kotus.fi/aineistot/vks_sahkoinenaineisto. shtml).

Electronic corpus for Finnish dialects. This collection is based on a multimillion set of records for Finnish dialects. Part of the material intended for the Finnish language dialect dictionary has been translated into electronic form [19]. The institute also stores an archive of grammatical phenomena concerned with the structures of Finnish language dialects (about 1,100,000 words, 200 thousand sentences). This corpus was prepared in the Finnish language department at Turku University (http: //www.utu. fi/hum/suomi/kokoelma.htm#LA).

The toponymic bank: this consists of 250 thousand toponyms from 40 areas in Finland, with about 10% converted to computer format [20].

In addition to the language materials, the institute has available the auxiliary programs KWIC, AGREP, and so on.

English language section at Helsinki University
http://www.eng.helsinki.fi/mdex.htm

Members of this division at Helsinki University were among the first to begin working on corpuses of English-language texts. Although there is a considerable number of research groups concerned with similar operations, the vision has made a substantial contribution to English corpus linguistics.

The Helsinki Corpus of English Texts: Diachronic Part [21]

This is probably the most famous corpus on the history of English and was prepared by an international group linked by the division. The corpus includes texts in various styles: from the earliest documents in Old English (since 750) to 18th century texts (1710). The corpus began to be set up in 1984. Its general volume is more than 1.5 million words. It does not contain linguistic annotation, although it includes 25 parameters of culture history annotation (author's name, date of creation, availability of foreign original, age, gender, social position of author, and so on). At present, various parts of the corpus serve as basis for an entire series of international projects, in which it is planned to perform multiaspect annotation, including linguistic [22].

The Helsinki Corpus of English Texts: the Dialect Corpus (HD) [23]

The Helsinki corpus of English dialects contains interpreted records from dialect studies made in 1970-1980 by Orton's field research method [24].

The corpus contains records of total volume 800 thousand words divided into several subcorpuses. The corpus is not annotated and in the opinion of its originators is suitable primarily for morphosyntactic purposes, and to a smaller extent for lexicological and phonetic research.

The Corpus of Early English Correspondence (CEEC) [25, 26, 27]

The original material was prepared in 1993-2000 in the English language section and includes texts from private letters of total volume more than 2.7 million words, which relate to the period from 1417 to 1681. The corpus has now been expanded with a subordinate body of texts from the later decades of the 17th century and the 18th century. The design principles have been based on the project "Sociolinguistics and language history", so the annotation corresponds mainly to those requirements (author's name, title, relation between author and recipient, and so on). The corpus is distributed on compact disks.

The Corpus of Early English Medical Writing [28]

This corpus covers 1375-1750 and has a total volume of about 1.5 million words. The contents range widely from theoretical studies to popular medical handbooks and guideline texts on medical topics. The work on the corpus has not yet been completed. It is planned to include morphological and syntactic annotation.

The Helsinki Corpus of Older Scots (HCOS) [29]

This unannotated corpus covers the period from 1450 to 1700. The texts are various styles of prose: acts of parliament, diaries, biography, pamphlets, scientific works, textbooks, letters, and so on. Volume about 830 thousand words.

Research Center for Languages of the Volga-Kama Area (Turku University) [30] and http://www.utu.fi/hum/sgr/VolgaPalvEngL.htm

The research group at Turku is concerned primarily with the Finno-Ugric languages in that region, namely Mordov, Marian, and Udmurt, but in connection with research on language contacts in that region, it covers also languages of the Turkic family (Chuvash, Tatar, Bashkir). A separate task of the group is the preparation of language corpuses. Some of them have been prepared, and publication on the Internet is planned.

MORMULA corpus: this consists of texts in the Mordov languages (Erzya and Moksha), total volume 200 thousand words, with morphological annotation.

MARKO corpus: this consists of unannotated texts from the Marian languages (mountain Marian and plains Marian), total volume about 1 million words.

ONCHYKO corpus: this consists of unannotated materials from a journal in plains Marian "Onchyko" for 1996-1999.

Erzya-Mordov corpus: this consists of literary texts and includes 16 transliterated texts, together with the New Testament in the Erzya-Mordov language (translated in the 19th century).

Individual projects

The Neo-Assyrian Text Corpus Project
http://www.helsinki.fi/science/saa/cna.html

There is active participation by the division of African and Asiatic research at Helsinki University in this international project, which is designed to collect published and unpublished neo-Assyrian texts and present them as an electronic database followed by publication of hard copies of the texts (SAA series, State Archives of Assyria). Unfortunately, we do not have information on the number of words in the corpus, and it is known only that the text part of the corpus in ANSII encoding is 3,358,547 bytes.

The Corpus of 19th-century English (CONCE) [31]

This is a joint project by the Department of English Philology at Tampere University and the English Language Division at Uppsala University (Sweden). It is a carefully balanced selection of texts in three separate parts (1800-1830, 1830-1870, and 1870-1900) in each of which there are set proportions of seven styles of text: official debates in parliament, court records, theatrical matters, news, personal letters, historical monographs, and scientific monographs in the area of natural or social sciences. The total volume is about 1 million words. The corpus is at present not accessible for public use.

Suomen Kansan Vanhat Runot (SKVR)
http;//www.oszk.hu/uidonsag/finn/eng/klemettinen.html

The Finnish Literature Society possesses one of the largest folklore archives in the world. Part of the materials written over centuries have been translated into digital form within the "Suomen kansan vanhat runot, SKVR" project ("Ancient tales of the Finnish people"). The project is being conducted in collaboration with researchers from Tartu (Estonia). It is proposed to prepare published and unpublished materials in the XML format.

It is planned in that way to prepare up to 150 thousand folklore texts, intended primarily for folklore researchers and literary studies.

The Electronic Corpus of Ingrian Finnish http://helmer. hit. uib. no/Ingrisk/ingrian. html

This comparatively small corpus represents audio recordings and interpretations of interviews with Ingermanland Finns. The field studies were conducted by experts from Joensuu [32]. The electronic corpus is a joint Norwegian-Finnish project.

Corpus Cyrillo-Methodianum Helsingiense http://www.slav. helsinki.fi/ccmh/

The electronic corpus involves translating into electronic form some ancient Old Slavonic texts. The work is being conducted at the Division of Slavonic and Baltic Languages and Literature at Helsinki University. The texts are presented in Latin transliteration with the addition of several special signs for abandoned letters and superscripts and so on. At present, the texts of Marian Evangelism, Suprasal manuscripts, and writings of Cyril and Methodius have been published on the Internet. It is planned to publish the texts of other ancient Old Slavonic manuscripts.

The Tampere Bilingual Corpus of Finnish and English [33]

This corpus of parallel Finnish and English texts has been set up by the translation section at Tampere University some considerable while ago. The basis of it is provided by the text of a novel by the Finnish writer Miki Valtari "Sinuhe egyptiläinen" and its translation into English. Tools and methods have been developed for setting up parallel texts that enable one to extend the corpus, to which have been added eight literary products (Finnish originals and English translations). At present it is planned to expand the corpus by adding literary and other texts. The corpus does not have linguistic annotation.

The PARRUS parallel Russian-Finnish corpus of literary products [34]

The basic idea of this project, which is being developed in the translation division at Tampere University, is to set up a two-language corpus of parallel texts in Russian and Finnish. It is concerned mainly with the texts of classical Russian literature and their translation into Finnish. The volume of the corpus is about 5 million text forms. The creators consider that this unsymmetrical corpus will support research on translation styles, the individual styles of the translators, and so on.

HANCO annotated Russian language corpus
http://www.slav.helsinki.fi/hanko

This corpus is now being prepared in the division of Slavonic and Baltic languages and literature at Helsinki University. In the first stage, it is planned to set up a database containing morphological and syntactic information. The material is provided by journal articles of total volume about 100 thousand text forms. The comparatively small volume of the corpus is compensated for by a branched multilevel annotation system, which involves not only automatic processing but also manual.

REFERENCES

1. M. Miettinen (editor), Kieliteknologia Suomessa, Helsinki, 1998.
2. C. Henriksen, E. Hovdhaugen, F. Karlsson, and B. Siguurd (editors), The History of Linguistics in the Nordic Countries, Helsinki, 2000.
3. K. Koskenniemi, "Tietokonelingvistiikan vaihteet Suomessa", Tietoyhteys, no. 3, pp. 7-8, 1997.
4. M. Miettinen, "Kieliteknologia tarvitsee Kielipankkia", Tietoyhteys, no. 3, pp. 5-6, 1997.
5. M. Lehtinen, "Tietokoneet sanakirjatyössä", Tietoyhteys, no. 3, pp. 9-10, 1997.
6. M. Lehtinen, P. Karvonen, and T. Rahikainen, Raportti tekstikorpusten koostamisperiaatteista ja nykysuomen tekstiaineistojen tarpeellisuudesta Kotimaisten kielten tutkimuskeskuksessa, Helsinki, 1995.

7. XXIX Kielitieteen päivät, Helsinki, 2002.
8. A. Arppe, No Single Path: Finnish Lessons in the Commercialization of Language Engineering Research, 2002. Access on address: http://www.esc.fi/euromap/artikkelit/arppe.phtml.en
9. T. McEnery and A. Wilson, Corpus Linguistics, Edinburgh, 1996.
10. M. Miettinen, Kielipankinasiakkaanopas, Espoo, 2000. Accessonaddress:http://www. esc. fi/kielipankki/julkaisut/opas/index .phtml
11. M. Grönroos, The Lexical Database of the Bank of Finnish, Helsinki, 2000.
12. P. Saukkonen (editor), Oulun korpus 1960-luvun suomen yleiskielen tutkimusmateriaali, Oulu, 1982.
13. A. Hakulinen, F. Karlsson, and M. Vilkuna, Suomen tekstilauseiden piirteitä, kvantitatiivinen tutkimus, Helsinki, 1980. [=Publications 6, Department of General Linguistics].
14. K. Koskenniemi, Two-Level F hrphology A General Computational Model for Word-Form Recognition and Production, Helsinki, 1983. [= Publications 11, Department of General Linguistics].
15. F. Karlsson, A. Voutilainen, J. Heikkilä, and A. Anttila (editors), Constraint Grammar: A Language-Independent Framework for Parsing Unrestricted Text, Mouton de Gruyter, Berlin, 1995.
16. P. Suihkonen, Documentation of the Computer Corpora of the Uralic Languages at the University of Helsinki, Department of General Linguistics, Helsinki, 1998. [=Technical Reports TR-2].
17. L. Viikki, RUSTWOL: A System for Automatic Recognition of Russian Words, 1997. Access on address: www.lingsoft.fi/doc/rustwol/rustwol.txt
18. Suomen kielen perussanakirja, Helsinki, 1990-1994, I-III.
19. Suomen murteiden sanakirja, Helsinki, 1985.
20. R. Miikkulainen, "The Database of Finnish Toponyms", Proceedings of the XIXth International Congress of Onomastic Sciences, Aberdeen, August 4-11, 1996, vol. 2, pp. 248-255, Aberdeen, 1998. Access on address: http://www.kotus.fi/aineistot/nimiarkisto/paikannimet/toponyms.shtml
21. M. Kytö , Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts, 3rd edition, Helsinki, 1996. Available at http://khnt.hit.uib.no/icame/manuals/HC/index.htm
22. M. Kytö and M. Rissanen, "English historical corpora: Report on developments in 1999", ICAME Journal, Computers in English Linguistics, no. 24, pp. 159-175, 2000.
23. K. Peitsara and A. Vasko, "The Helsinki dialect corpus: Characteristics of speech and aspects of variation", Helsinki English Studies, The Electronic Journal of the Department of English at the University of Helsinki, vol. 2, 2002. Access on address: http://www.eng .helsinki.fi/hes/Corpora/extending_the_corpus.htm
24. H. Orton, Survey of English Dialects, Leeds, 1962.
25. T. Nevalainen and H. Raumolin-Brunberg, "Sociolinguistics and language history: The Helsinki Corpus of Early English Correspondence, Hermes", Journal of Linguistics, vol. 13, pp. 135-143, 1994.
26. A. Nurmi (editor), Manual for the Corpus of Early English Correspondence Sampler CEECS, Helsinki, 1998. Access on address: http://www.hit.uib.no/icame/ceecs/index.htm
27. M. Laitinen, "Extending the corpus of Early English correspondence to the 18th century", Helsinki English Studies, The Electronic Journal of the Department of English at the University of Helsinki, vol. 2, 2002.
Access on address: http: //www. eng .helsinki. fi/hes/Corpora/extending the corpus .htm
28. L Taavitsainen and P. Pahta, "Corpus of Early English medical writing, 1375-1750", ICAME Journal, Computers in English Linguistics, vol. 21, pp. 71-79, 1997.
29. A. Meurman-Solin, "A new tool: The Helsinki corpus of Older Scots (1450-1700)", ICAME Journal, Computers in English Linguistics, vol. 19, pp. 49-63, 1995.
30. A. Moisio and J. Luutonen, Turun yliopiston volgakielten Korpukset, XXIII Kielitieteen päivät, 123, Helsinki, 1996.
31.M.Kytö, J.Rudanko, and E. Smitterberg, "Building a bridge between the present and the past: : corpus of 19th-century English, ICAME Journal, Computers in English Linguistics, vol. 24, pp. 85-97, 2000.
32. W. R. Cooper, "The Tampere bilingual corpus of Finnish and English: Development and applications, compare or contrast?" Current Issues in Crosslanguage Research, Tampere, 1998.
33. I. Savijärvi, Western Ingria: Where Languages and Dialects Meet. Access on address: http: //helmer .hit .uib. no/Ingrisk/western .html
34. M. Mikhailov and H. Tommola, "Compiling parallel text corpus: Towards automation of routine procedures", International Journal of Corpus Linguistics, vol. 6, pp. 69-77, 2001.