Twenty Years of Old Literary Finnish – from Manual to Computer-Assisted Research[1]

Until fairly recently, research on old Finnish was either based on original texts, or on facsimiles, microfilms and similar sorts of copies in the so-called Old Fennica collection at the University Library of Helsinki. Since the early 1990s, KOTUS, i.e. the Research Institute for the Languages of Finland in Helsinki (*http://www.kotus.fi/*)[2] has been transferring its materials to computer files, thereby making computer-assisted research possible. As of March 2003, the computerized corpora available for researchers other than the Institute's own staff comprise some 17 million Finnish words.[3] The commands of the Institute's Unix system can be used for processing data in the corpora, but several additional search programmes are also available for researchers; *agrep* is used for searching approximate character strings, *kwic* for creating KWIC indexes, and *segrep* for searching character strings in defined text structures. Today the Centre uses the Unix system exclusively, having recently discarded the Open VMS system that was previously used parallel to Unix.

The Institute's Finnish corpus is divided into four sub-groups;

(1) *Old Literary Finnish* (1540—1810), approx. 2.8 million words.

(2) *19th Century Finnish* (1810—1900), approx. 4 million words.[4]

(3) *Modern Finnish* consists of **dictionaries** (a *Dictionary of Modern Finnish* from 1966, completed with a glossary of neologisms from the 50s, 60s and 70s, as well as a *Basic Dictionary of Finnish* in three parts, published in 1990, 1992 and 1994, respectively); a **text bank** of Finnish from the1990s, and a **data base** of some 100,000 word entries with an annual increase of approx. 8000.

(4) A corpus of *Finnish dialects* together with *Finnish nomenclature*, i.e. Christian and family names as well as geographical names.

---

[1] The present article, under the title *Computerized Corpora – Shortcuts to Research into Old Literary Finnish*, appeared in the Journal of Finnish Studies (Volume 3, No. 2), University of Toronto, Canada, 1999. It has now been updated, revised and somewhat extended. When updating the article I was given advice and statistical information by Tarmo Rahikainen, the IT Manager of the Research Institute for the Languages of Finland.

[2] In the late 1990s the name of the website was changed from *domlang.fi* [i.e. domestic languages] to *kotus.fi*.

[3] Today the Finnish ICT center for science, CSC, owned by the Ministry of Education, provides universities and research institutes with Finland's widest selection of scientific software and databases. Since CSC offers access to the collection of Finnish texts of the Language Bank of Finland (approx. 180 million words) as well as to the so-called Oulu corpus, KOTUS only offers access to these corpora to its own staff, not to researchers at large.

[4] That is, more than 3.3 million text words and nearly 0.7 million dictionary words.

The so-called Oulu corpus, which contains nearly half a million words of Finnish text from the 1960s, was SGML-coded and included in the Institute's Modern Finnish materials (No. 3 above). The Institute's corpus of Finnish dialects (No. 4 above) is based on some eight million pages of (handwritten) notes. It also includes the materials of the Finnish Syntax Archives, some 1.1 million words, which were SGML-coded before they were added to the Institute's computer files. A major dictionary of Finnish dialects (*Suomen murteiden sanakirja*) is being produced on the basis of these materials. This dictionary will be accessible to users not only in a printed version but in a computerized version as well. When completed, this dictionary will consist of twenty parts (approx. 18,000 pages) with totally 300,000 to 350,000 different words. Part 1 was published in 1985, and a separate introductory book followed in 1989. So far six parts have been completed (alphabetically from *a* to *keynätä*), with the following parts appearing at approximately three-year intervals. A joint project of the Research Institute, the Department of Finnish at the University of Helsinki, and the Finnish Academy, namely, a *Descriptive Grammar of Finnish*, was initiated in 1995, and the work in manuscript form is expected to be finished in the current year (2003).

COMPUTERIZED CORPORA OF OLD AND EARLY MODERN FINNISH

The old Finnish materials are stored in SEN files (e.g., *ahf1700.sen* = 18th century Acta Historica Fennica), which means that every sentence or every piece of text ending in a period will appear as one computer line. Each line begins with a location code that indicates, among other things, the title of the work, name of the author, year of the publication, and the original page number. Anyone wishing to see what is stored in the Institute's computers may open the link to "Aineistot ja arkistot" ('materials and archives') on the Institute 's home page. The links there include one to "Sähköiset aineistot" ('computerized materials'), and this link will provide search instructions as well as a list of corpora.

The corpus of *Old Literary Finnish* (1540—1810) contains some 2.8 million words or more than 20 million characters. The first texts to be computerized were the works in three volumes of the first scholar to write in Finnish, Mikael Agricola (approx. 1543—1552), altogether almost half a million words. The list of old Finnish materials comprises some twenty titles, among them a book of hymns, a book of mediaeval Latin songs, the first Finnish Bible translation, translations

of the Swedish law book, statutes and old historical documents, the annual almanacs, the very first newspaper in Finnish, four works by Christfrid Ganander in addition to his extensive 18th century dictionary, and a book on biology by Johan Frosterus. See the list of *Computerized Materials of Old Literary Finnish* on the web pages of KOTUS: (http://www.kotus.fi/aineistot/vks_sahkoinenaineisto.shtml)

The Institute has used the corpus of Old Literary Finnish to produce computer-based dictionaries and various indices, including, among others

- *Index Agricolaensis* I & II;
- *Dictionary of Old Literary Finnish* 1 (a—i) & 2 (j—k), with part 3 expected to appear in 2004 and parts 4 to 6 by the end of the 2020s;
- *Christfrid Ganander's Dictionary* from 1787 (originally a manuscript) with an index.

The materials of *Early Modern Finnish* (1810—1870) comprise part of the Institute's 19th century corpus, which has a scope of some 4 million text words. The collection of 19th Century Finnish (approx. 1810—1900) is a fairly recent project at the Research Institute. This project was included in the Institute's program in 1992 for the purpose of computerizing the literary materials covering the period 1811—1880 (see Punttila 1993 & 1995), which had not been covered by the dictionaries produced at the Institute; the *Dictionary of Old Literary Finnish* did not go beyond the year 1810, whereas the *Dictionary of Modern Finnish* had the year 1880 as a starting point. The number of titles of the computerized materials of 19th century Finnish increased from some twenty in 1997 to nearly one hundred in 2002. The project also envisioned a plan for a 19th century dictionary in computerized form (SGML-coded), thus enabling research on Finnish morphology and syntax within the above-mentioned period.

The project for computerizing 19th century literature has been supported by several partner institutions at Finnish universities and at the University of Tartu, Estonia. When completed, the 19th century Finnish corpus will include a balanced and representative selection of texts from different fields, such as biology, economics, fiction, history, linguistics, mathematics, medicine, sports, and technology. The corpus index, which was updated in 2001, comprises texts by more than 50 authors[5] as well as articles covering various genres, from statutes to newspapers and periodicals, from grammars, school books and dictionaries to the collected works of such authors as Aleksis Kivi and Elias Lönnrot. Authors with several items in the list – besides Aleksis Kivi and Elias Lönnrot – are Pietari Hannikainen (dramas and drama translations), Jacob

[5] In 1999 this corpus only contained texts by fifteen authors.

Juteini (devotional prose), Agaton Meurman (history and politics), Henrik Renqvist (religious and devotional prose), Samuel Roos (devotional prose), Carl Axel Gottlund (a reader and a book of poems), E. Salmelainen (sagas and legends), and Zachris Topelius (folk poetry).

The following issues of newspapers and periodicals have been transferred to computer files: Kirjallinen Kuukauslehti (file name: *kkl*) 1867 and 6/1869; Keski-Suomi (*keskisuomi*) 1/1871; Mehiläinen (*mehilainen*) 1859; Oulun Viikko-Sanomia (*ovs*) 1829, 1830, 1831, 1832; Sanan Saattaja Viipurista (*ssv*) 1833, 1841; Suometar (*smtr*) 1847, 1848; and Turun Viikko-Sanomat (*tvs*) 1820, 1823, 1830. The 19th century dictionaries that have been computerized are those of Gustavus Renvall 1823—1826, parts 1 & 2 (Finnish-Latin-German), Carl Helenius 1838 (Swedish-Finnish), Elias Lönnrot 1847 (Swedish-Finnish-German), and a Finnish Parlor ('Finsk parlör') from 1860. The statute corpus (*/korpus/1800/asetus/)* contains some 500 titles, i.e. the annual statutes issued in the years 1811—1899. For details, see the list of *Computerized Materials of 19th Century Literary Finnish* on the web pages of KOTUS: (http://www.kotus.fi/aineistot/1800/1800_sahkoisetaineistot.shtml)

A report on the Research Institute's computerized text corpora (Lehtinen & al. 1995) stressed the need for a more substantial corpus of modern Finnish texts. Following this, an EU-financed 'Parole project', in other words a text bank of Finnish, was initiated at the Institute in 1996 in cooperation with the University of Helsinki's Department of Linguistics. At the end of the two-year project in 1998, over 21 million words of modern Finnish text had been SGML-coded and made available for researchers. This was just the beginning, especially when considering such foreign collections as the Bank of English (200 million words), or the British National Corpus (100 million words). The 'Parole corpus' which, besides Finnish, includes Swedish texts from Sweden and Finland, has been extended through several other projects, and its present scope is some 180 million words. Today the entire corpus, known as the collection of Finnish texts of the Language Bank of Finland, is accessible to users on the web pages of CSC (see footnote No. 3). Another Nordic text bank, the Språkbanken of the University of Gothenburg (see Lehtinen & al. 1995), has been created in Sweden (the Swedish 'Parole corpus' is part of it).

When the Research Institute of Helsinki joined the Finnish university network Funet[6] in 1992, Internet users gained access to the Institute's corpora. Users naturally require a password.

[6] Funet data communication links are today included in CSC's services for researchers.

Despite the option of using the Open VMS system, I preferred to use the Unix system, and preferably its *kwic* command, when working with the Institute's corpora.

FROM MANUAL TO COMPUTER-ASSISTED RESEARCH

Since the early 1980s my principal field of research has been the syntax of Old Literary Finnish. The first result of this work was my doctoral thesis of 1983, entitled *Satsmotsvarigheter i finsk prosa under 1600-talet* ('Abbreviated Clauses in 17th Century Finnish Prose'). The extensive corpus that I excerpted for the thesis (some 2.5 million words) needed to be handled manually; in those days there were no computerized corpora in existence for either Finnish or Swedish scholars. A few years later, in 1988, I was granted a scholarship by the Swedish Research Council for the Humanities and Social Sciences (Sw. Humanistisk-Samhällsvetenskapliga Forskningsrådet, abbr. HSFR)[7] for the project *Finsk syntax på 1600-talet* ('Finnish Syntax in the 17th Century'), which, to a great extent, was a continuation of my thesis work. This research project, which among other things resulted in the two monographs *Vanhan kirjasuomen teonnimistä ja teonnimirakenteista* ('On Deverbal Nouns and Deverbal Constructions in Old Literary Finnish', 91 pp.), 1990, and *Vanhan kirjasuomen nominaalirakenteista* ('On Non-Finite Constructions in Old Literary Finnish', 120 pp.), 1992, was also completed using the traditional manual technique (this research was finished at the end of 1993). The only difference when compared to the time of my thesis work was that I now used a computer and a word processing program.

It was not until my second research project was initiated in 1995 that I became familiar with computerized corpora. The Swedish Research Council was willing to finance the continuation of the 17th century project, i.e. *Studier i finsk syntax från 1600-talet till mitten av 1800-talet* ('Studies in Finnish Syntax from 17th Century to Mid-19th Century'). When applying for this grant in January 1995 I was aware of the fact that the Helsinki Research Institute for the Languages of Finland had been computerizing its corpora for some time, and that materials from the Old Literary Finnish and Early Modern Finnish periods were well represented. In this research project I intended to employ computerized corpora to the extent they were available for my purposes. Fortunately, there has been a steady increase in the bulk and variety of the Institute's computerized corpora from 1995 onwards.

[7] The Council has recently changed its name to the Swedish Research Council (Sw. Vetenskapsrådet, abbr. VR).

The new project was to begin where the previous one had left off, that is, at the turn of the 17th and 18th centuries. The period to be investigated covered nearly 150 years, with some of that time belonging to the period of Old Literary Finnish (until 1810), and some to the Early Modern Finnish period (from 1810 onwards). My criterion for selecting certain morpho-syntactic phenomena for analysis was that there was a clear difference between 17th century usage and modern usage. The main topics to be examined were as follows:

1. The *in* and *on* cases of the third infinitive, i.e. *luke/**ma/ssa*** '(is) reading', *luke/**ma/lla*** 'by reading'.
2. The absolute instructive (*kyynelissä silm/**in*** 'with tearful eyes'), and the absolute nominative (*silmä/**t** kyynelissä* 'with tearful eyes' or 'with one's eyes in tears).
3. The predicative in the partitive case (*Tämä on vet/**tä*** 'This is water'; *Ei se sitä laatu/**a** ole* 'It is not of that sort').
4. The (pragmatic) clitics *-hAn*, *-pA*, *-s*, *-stA* (e.g. *Tämä/**hän** ~ Tämä/**pä**(**s**) on hyvää* 'I say, this is nice').
5. Verbs of sense perception and their qualifiers, for example *kuulua* 'sound', *näkyä*, *näyttää* 'seem', *tuntua* 'feel', and several others (*Se näkyy minu**lle** kumma**ksi*** 'It seems strange to me' vs. Modern Finnish *Se näyttää minu**sta** kumma*[*llise*]***lta***).
6. The so-called agent participle or *mA* participle (e.g. *Jumalan luo/**ma*** 'created by God'), vs. deverbal nouns with the same suffix (e.g. *elä**mä***, *kuole**ma***).
7. The so-called quasi construction (e.g. *Hän on luke/**vi/na/an*** 'He pretends to read').
8. The use of the potential mood in poetry and prose (e.g. *Missä veljeni Abel **lienee***? 'I wonder where my brother Abel is'; *Oll**e**nko minä veljeni vartija*? 'Am I/Should I be my brother's guard').
9. Some modal constructions expressing necessity (e.g. *On teh**tävä*** 'You have to do', *Ei **tarvitse** tehdä* 'You need not do').
10. New types of deverbal nouns, formed by the suffixes *-ntA*, *-nti* and *-UU* (e.g. *laula**nta*** 'singing', *juo**nti*** 'drinking', *leikk**uu*** 'cutting').

From the very start of my second project I judged it wisest to combine the two research techniques, manual excerpting and searching in computer files. Thus, a representative selection of 18th century materials, religious and other texts, were manually excerpted. Following this, the same constructions or phenomena (see the topics above) were located in a number of

computerized materials: *ahf1700* (= Acta Historica Fennica 'Finnish historical documents'), *as1700* (= Asetus 'statutes'), *rwl1759* (= Ruotzin Waldacunnan laki 'law of Sweden'), *varia1700* ('various/miscellaneous texts'), *ssv1833* & *ssv1841*, *tvs1820* & *tvs1823*, and *smtr1847* & *smtr1848* (cf. the titles and file names in the passage on newspapers and periodicals above). File search was possible if the examined phenomenon had a distinct morphological form. On the other hand, certain syntactical aspects could only be identified by manual excerpting. An example of this would be the predicative in the partitive case, as search commands are unable to separate partitives with a predicative function from partitives with other functions because the text corpuses have not been coded for this purpose.

Many of the 17th century materials which were previously only accessible manually now exist in computer files. One example of this is the large corpus of the first Finnish Bible translation of 1642. In my latest research, I have utilized the Bible corpus to study the earlier phases of certain linguistic phenomena. This was the case when I examined the so-called quasi construction (Old Finnish *on teke/**wä/nä/ns***, Modern Finnish *on teke/**vi/nä/än*** 'pretends to do'). While data had been manually collected from texts dating from the period 1700—1850 or thereabouts, 17th century usage was examined by locating the same constructions in the computerized Bible text. For further data I also searched the Institute's 18th and early 19th century computerized materials. Since the quasi construction has a distinct morphological marker, it was possible to locate the appearances with a few search commands.

COMPUTERIZED MATERIALS — SHORTCUTS OR NOT?

A number of problems—sometimes minor, sometimes major—are involved with file searches. The nature of the linguistic features investigated determines whether search commands will be successful or not.

1. If the linguistic feature has a distinct morphological form, commands that use the specific morpheme(s) for the search will locate the entire set of occurrences in text corpora (e.g., the quasi construction, cases of the third infinitive, such clitics as *-hAn* or *–pA*, and deverbal suffixes such as *-ntA*, *-nti*, *-UU*).
2. A distinct morphological element may bring forth most occurrences of the examined construction, but, at the same time, a large amount of irrelevant information will be

received. The same morpheme—or more precisely, the same set of characters—may have several functions, and the search will produce hundreds of irrelevant words. This would be the case with the *mA* participle because the same suffix is used to form participles, infinitives and nouns. A search for the clitic *-s* would produce a myriad of *s*-forms, in which *s* would represent a word formation element, a case marker, a tense or a mood marker, and some other things as well.

3. In some cases a search for morpho-syntactic phenomena can only be conducted by way of vocabulary. "Absolute instructives" look like any other instructives and, furthermore, instructive forms can be identical with other case forms. Knowing that the constituent in the instructive case in the absolute instructive construction normally indicates body parts, a search can be conducted by using such nouns in the instructive case. Preferably, vocabulary can also be used in the search for qualifiers to verbs of sense perception; if one searches for certain verbs (= certain strings of characters), possible qualifiers will follow.

It is true that computerized text corpora offer shortcuts to research on many morpho-syntactic phenomena. However, when investigating purely syntactic or morphologically diffuse linguistic features, traditional (manual) research will have to be used or at least combined with the searches.

BIBLIOGRAPHY

FORSMAN SVENSSON, PIRKKO 1983. *Satsmotsvarigheter i finsk prosa under 1600-talet: participialkonstruktionen och därmed synonyma icke-finita uttryck i jämförelse med språkbruket före och efter 1600-talet*. SKST 388. Helsinki: Suomalaisen Kirjallisuuden Seura.

— 1990. Vanhan kirjasuomen teonnimistä ja teonnimirakenteista. *Stockholm Studies in Finnish Language and Literature* 5. Umeå: Umeå universitets tryckeri.

— 1992. Vanhan kirjasuomen nominaalirakenteista. *Stockholm Studies in Finnish Language and Literature* 8. Uppsala universitet: Reprocentralen HSC.

GANANDER, CHRISTFRID 1787. *Nytt Finskt Lexicon*. Alkuperäiskäsikirjoituksesta ja sen näköispainoksesta toimittanut Liisa Nuutinen. SKST 676. KKTK:n julkaisuja 95. Helsinki 1997.

Index Agricolaensis I-II. KKTK:n julkaisuja 11. Helsinki: Kotimaisten kielten tutkimuskeskus.

KKTK = Kotimaisten kielten tutkimuskeskus.

KOTUS = Kotimaisten kielten tutkimuskeskus 'Research Institute for the Languages of Finland' (http://www.kotus.fi).

LEHTINEN, MARJA & PIRJO KARVONEN & TARMO RAHIKAINEN 1995. *Tekstikorpukset. Raportti tekstikorpusten koostamisperiaatteista ja nykysuomen tekstiaineistojen tarpeellisuudesta Kotimaisten kielten tutkimuskeskuksessa*. Helsinki: Kotimaisten kielten tutkimuskeskus.

NUUTINEN, LIISA (ed.) 1997. *Christfrid Gananderin Nytt Finskt Lexicon. Hakemisto*. SKST 688, KKTK:n julkaisuja 100. Helsinki: Kotimaisten kielten tutkimuskeskus.

Nykysuomen sanakirja 1966. Helsinki: WSOY.

PUNTTILA, MATTI 1993. 1800-luvun kirjasuomen keruuhanke. *Kieliposti* 1/1993 s. 42—43.

— 1995. 1800-luvun kirjasuomen aineslähteistä ja niiden käytöstä. *Acta Universitatis Ouluensis* B 19 s. 207—214. Oulu.

SKST = Suomalaisen Kirjallisuuden Seuran Toimituksia.

Suomen kielen perussanakirja 1—3. KKTK:n julkaisuja 55. Helsinki: Valtion painatuskeskus 1990, 1992, 1994.

Suomen murteiden sanakirja 1—5. KKTK:n julkaisuja 36. Helsinki: Valtion painatuskeskus 1985, 1988, 1992, 1994, 1997.

Vanhan kirjasuomen sanakirja I—II. KKTK:n julkaisuja 33. Toim. LAHJA IRENE HELLEMAA, ANJA JUSSILA, ESKO KOIVUSALO & RIITTA PALKKI. Helsinki: Painatuskeskus 1985, 1994.