A Corpus of Written Finnish Romani Texts

Lars Borin

Department of Linguistics, Uppsala University
Box 527, SE–751 20 Uppsala, SWEDEN
Lars.Borin@ling.uu.se

Abstract

Finnish Romani is a language with a fairly recent written tradition; for all practical purposes it is a 20th century phenomenon. An official orthography was created in 1971, and it is mostly from the 1970’s onwards that we see texts of the kind which we normally associate with a written language variety. The text corpus described here is being compiled to support an ongoing investigation into the effects of language contact on Finnish Romani.

1. Introduction

1.1. Finnish Romani

Finnish Romani is the language of the Finnish Roma (or Gypsies), a minority now numbering on the order of ten thousand persons. The Finnish Roma have been living in Finland and Sweden since the 16th century, and at present about two thirds of the population live in Finland and the remaining third in Sweden.

The language itself belongs to the so-called Northern Romani dialect branch, its nearest relatives being the British, Sinti–Manush and Polish–North Russian–Baltic Romani languages/dialects (Bakker and Matras, 1997).

The sociolinguistic situation of Finnish Romani is that it is both a minority language and a language which has been yielding to Finnish for more than a century. It is probably safe to say that there are no, or extremely few, small children who speak the language (see figure 1), so if we were to take this commonly used indicator of the health of a language (see, e.g., Krauss, 1996) at face value, we would conclude that Finnish Romani could soon be counted among the extinct languages. This would be a rash assumption, however, as this state of affairs has obtained for more than fifty years, but the language is still used and very much alive.

This is partly the result of a conscious revival effort of fairly recent date, which has led to the introduction of Romani language instruction for Finnish Roma children in primary schools in Finland and Sweden. More than this, however, it reflects a trait which sets (modern) Finnish Romani apart from most other languages: it is not learned primarily in childhood, but gradually as Roma children are introduced into the life and activities of adulthood, where the language is used as a secret language (Valtonen, 1968, 241ff).

Finland has signed and ratified the European charter for regional or minority languages (Council of Europe, 1992) for Finnish Romani. This means that the language now enjoys certain rights in Finnish law, and also that more resources are allocated to Romani linguistic research and development of language resources, through the Research Institute for the Languages of Finland.¹

Finnish Romani is a language with a fairly recent written tradition; for all practical purposes it is a 20th century phenomenon. An official orthography was created in 1971 (Ortografiakomitea, 1971), and it is mostly from the 1970’s onwards that we see texts of the kind which we normally associate with a written language variety.

1.2. Corpora for minority languages

Corpora, especially parallel corpora, have possibly a more important role to play for minority languages than for well-established majority languages, particularly in terminology development and lexicography, but also as general resources in research on bilingualism and language contact (see Trosterud, to appear), and for the creation of computerized language instruction tools, and translation and writing aids.

¹http://www.domlang.fi/
2. The text corpus and its use

2.1. Composition of the corpus

The text corpus described here is being compiled to support an ongoing investigation into the effects of language contact (with Finnish) on Finnish Romani.

The corpus itself is a so-called convenience sample, i.e. it consists of those texts which have been available to the author, but nevertheless it also represents a significant fraction of the entire written Finnish Romani production.

The total size of the corpus is about 110,000 words, and there are mainly four kinds of texts in it:

1. Original articles by various authors from the periodical Romano Boodos (about 170 articles, 75,000 words; see figure 1 for an example)

2. Two language textbooks, (Koivisto, 1984; Koivisto, 1987) (9,500 words)

3. A translation into Finnish Romani of the Gospel according to John, made by Viljo Koivisto (Suomen Pipliesauru, 1971) (18,000 words; see figures 3 and 4)


The corpus also contains the Estonian linguist Ariste’s (Ariste, 1938) transcriptions of his fieldwork interviews—i.e., oral texts—made in Finland in the 1920’s (950 words), and a Romani-Finnish wordlist with 3,000 entries (from Ortografiakomitea, 1971).

Among the more sizeable written text materials missing from the corpus are:

1. Newer teaching materials, such as the textbooks by Vuolasranta (Vuolasranta, 1995) and Hedman (Hedman, 1996)

2. The translations of the Gospels according to Mark (by Valtonen) and Luke (by Hedman)

3. Most of the existing dictionaries, such as Thesleff’s Romani-German dictionary (Thesleff, 1901), Valtonen’s Romani-Finnish etymological dictionary (Valtonen, 1972) and Koivisto’s Romani-Finnish-English dictionary (Koivisto, 1994).

Also missing are most oral texts, such as Valtonen’s (Valtonen, 1968) and Vuorela’s (Vuorela and Borin, 1998) fieldwork transcriptions.

Figure 1: An article in Romano Boodos No. 2, 1992.

2.2. Use of the corpus

As mentioned above, the corpus is intended as a source of data for linguistic investigations of contact phenomena in Finnish Romani. Ongoing investigations using these texts are concerned with
• the possible emergence of a new definite article (the original Romani articles have disappeared) on the pattern of Finnish *se* (see Laury, 1997)

• the development of an infinitive, replacing an earlier use of finite forms in dependent clauses, à la Balkan languages (see Boretzky, 1996)

• the transformation of the possessive form from a denominal adjective into a nominal case form (genitive; see Koptjevskaja-Tamm, forthcoming)

• general Finnish influence in lexicon, phraseology and syntax, e.g., the replacement of prepositions governing the nominative, oblique or locative, by postpositions governing the (new) genitive.

At first, the text corpus was processed by fairly simple means. The texts have been manually compiled and typed in over a period of several years, into text files identified by a simple ID expression at the beginning of each file. Thus, the text in figure 1 is identified as #RB:1992/02[AB1=O]. The parts of the ID expression identify the source (RB – Romano Boo- dos; 1992/02 – No. 2, 1992), the author (AB – Armas Baltzar, 1 – the first contribution by this author in the publication in question), and, finally, the text type (O – original written text, as opposed to a translated (T) or a transcribed spoken (S) text).

For organizing and processing the corpus, SIL’s Shoebox program (Summer Institute of Linguistics, 1998) and Perl scripts written by the author have turned out to be sufficient for many purposes. A Shoebox database was automatically created with Perl scripts from the texts in the format just described, where each database record consists of a ‘sentence’ (identified with a fairly simple tokenizing algorithm). ID fields are automatically generated from the text ID, plus a sentence number.

The filtering facilities of Shoebox have been used to extract database subsets, records containing particular interesting words or constructions. In figure 2, we see some records containing the word *dola* ‘that’ (oblique or plural), which may be developing into a new definite article in Finnish Romani (see above), although this hypothesis still awaits confirmation.

Recently, we have decided to take advantage of the fact that our department has a strong re-
search tradition in corpus linguistics, especially in the area of parallel corpora, where no less than two groups of researchers are working on complementary but related parallel corpus projects (Borin, to appear; Sågvall Hein, to appear).  

Thus, we have started to apply the sentence and word alignment tools developed in these projects (see Tiedemann, 1998; Tiedemann, to appear) to the problem of making explicit the parallel structure in the translated texts, and later, we will be able to use the tools for word alignment evaluation (Merkel, to appear) and sentence and word alignment browsing (Olsson and Borin, to appear), for which activities these projects have developed software tools.

We have aligned the Finnish Romani Bible text on the sentence and word levels with its Finnish counterpart, thus making it easier to investigate contact phenomena in the lexical, phraseological and syntactic domains. Figure 3 shows part of a sentence aligned parallel text (the beginning of John 1), and in figure 4, we see how an incremental word alignment program aligns (some of) the words in the same text portion. As a concrete example of a contact phenomenon in this text portion, we may consider the expression "Deevelesko neere ‘with God’ in the first alignment unit. It should be analyzed

Deevel=es -ko neere
Jumala -n luona
God GEN with

i.e., with completely parallel structures in the two languages.

3. Further work

The availability of sentence and word alignment tools opens many possibilities for further investigation of contact phenomena in Finnish Romani, e.g.:

We know from our work with other language pairs that the precision 3 of word alignment is somewhere between 25 and 40 percent, depending on the language pair and the text type. We plan to investigate this for the pair Finnish Romani – Finnish, and we entertain the initial hypothesis that the precision will turn out to be high, because of a high degree of contact-induced structural and lexical correspondence between the two (totally unrelated) languages.

4. Conclusions

We have shown some examples of how our corpus of Finnish Romani—together with parallel material in Finnish—supports a variety of investigations into the effects of language contact on the language. As we mentioned earlier, such a corpus also forms a useful resource in work with language revitalization, maintenance, planning and standardization.

Minority languages—and so-called lesser used languages in general—may actually get a free ride from the work done on major languages, because much of that work is aimed at developing methods for automatic acquisition of linguistic knowledge from raw text, even in small quantities. These methods may then be used on text in a small language, to bootstrap linguistically structured resources which might not have come into being otherwise, because of lack of people with linguistic training. This means that even fairly small corpora—both monolingual and parallel—of minority and lesser used languages are well worth compiling.

5. Acknowledgements

Part of the research described here was carried out within the ETAP project, funded by the Bank of Sweden Tercentenary Foundation as part of the research programme Translation and Interpreting — a Meeting between Languages and Cultures (see http://www.translation.su.se).

My late wife Katri Vuorela first introduced me to the fascinating world of the Finnish Roma and their language, which was to have been the topic of her Ph.D. thesis (see Vuorela and Borin, 1998). Leif-Jöran Olsson in the ETAP project saw to it that the sentence and word alignment of

---

2 See also http://stp.ling.uu.se/etap/ and http://stp.ling.uu.se/~corpora/plug/.

3 The word alignment precision is the number of alignments found divided by the total number of possible alignments in the text pair.

4 Romani belongs to the Indo-Aryan subbranch of the Indo-Iranian branch of the Indo-European language family, while Finnish belongs to the Baltic-Finnic subbranch of the Fenno-Ugrian branch of the Uralic language family.
the Bible text went smoothly. The word alignment program used was conceived and developed by Jörg Tiedemann in the PLUG project (see http://stp.ling.uu.se/~corpora/plug/). Finnish Romani is still considered a secret language by many Finnish Roma (cf. Valtonen, 1968, 241ff), which means that the corpus described here cannot at the moment be made freely available to the public. Please contact the author for more details.

6. References
Council of Europe, 1992. European


Figure 4: Romani–Finnish word alignment of the beginning of John 1 (for English translations, see the appendix).
Appendix

Translations for figure 2

The translations of *dola* are boxed in.

AR: 1940/00 [PA1=S]0048: That is a charge of two hundred for rent, for the apartment.

AR: 1940/00 [PA2=S]0028: There comes a (non-Rom) woman and washes the child.

DR: 1982/11 [VK1=O]0004: First I heated water in the sauna kettle, and then when the water was hot and ready, I took the water out of that kettle and put it in the washing tub and started to wash the clothes.

DR: 1982/17 [VK1=O]0004: Those things in the museum were old-fashioned things.

JE: 1971/04 [VK1=T]0060: (John 4:45) When he arrived in Galilee, the Galileans welcomed him. They had seen all that he had done in Jerusalem at the Passover Feast, for they also had been there.

RB: 1992/01 [VK3=O]0016: Because of that, we need to learn more about these things.

RB: 1992/01 [VK3=O]0030: Old people have taught their children about the matters and things which they themselves have thought that they would be good for children and young people.

RB: 1992/01 [VK3=O]0041: And the people who themselves have received a good education, it would be good for them to start teaching their own people.

RB: 1992/01 [VK3=O]0051: It would be very good too, if in the Roma children’s homes there would be many such people who could teach the Roma children more about such subjects that they need in school.

Translation of the text in figures 3 and 4

1. In the beginning was the Word, and the Word was with God, and the Word was God.

2. He was with God in the beginning.

3. Through him all things were made; without him nothing was made that has been made.

4. In him was life, and that life was the light of men.

5. The light shines in the darkness, but the darkness has not understood it.

6. There came a man who was sent from God; his name was John.

7. He came as a witness to testify concerning that light, so that through him all men might believe.