Spoken to Spoken vs. Spoken to Written: Corpus Approach to Exploring Interpreting and Subtitling

Mikhail Mikhailov, Hannu Tommola, and Nina Isolahti

Abstract—The need for corpora of interpreting discourse in translation studies is gradually increasing. The research of AV translation is another rapidly developing sphere, thus corpora of subtitling and dubbing would also be quite useful. The main reason of the lack in such resources is the difficulty of obtaining data and the inevitability of manual data input. An interpreting corpus would be a collection of transcripts of speech in two or more languages with part of the transcripts aligned. The subtitling and dubbing corpora can be designed using the same principles. The structure of the corpus should reflect the polyphonic nature of the data. Thus, markup becomes extremely important in these types of corpora. The research presented in this paper deals with corpora of Finnish-Russian interpreting discourse and subtitling. The software package developed for processing of the corpora includes routines specially written for studying speech transcripts rather than written text. For example, speaker statistics function calculates number of words, number of pauses, their duration, average speech tempo of a certain speaker.

Index terms—Interpreting, subtitling, corpora, Russian language, Finnish language.

I. INTRODUCTION

COMPILING **written** text corpora has become a relatively easy technical task in the last decades. Some of published texts are ready available in digital form, other can be digitized with the help of scanning and OCR software. Plenty of texts of different genres written in all imaginable languages are being accumulated on the web. It is even possible to collect so called web-corpora in automated mode from the Internet (see works by Adam Kilgariff, William Fletcher, Marco Baroni, e.g. [1]). Text corpora exceeding 100 millions running words in size are quite common today[1].

As regards compiling **spoken** corpora, it remains hard, time-consuming, expensive and extremely slow work. As opposed to written resources, not many ready-made transcripts of spoken language are available (e.g. speeches of politicians, TV interviews, etc.), and most of those are adaptations of oral speech into written form and have to be matched with the recordings. As opposed to written resources, transcripts or

oral speech are not subject to amateurish collecting. Recording and transcribing of oral speech remain scholars' activity. Although the quality of sound recording and possibilities for data storage have greatly improved during the last decades, speech recognition technologies are still under development, error rate is considerably high [2]. The speech recognition systems are not being developed for converting spontaneous speech into textual form, but rather for the purposes of dictation. The commercial software is still quite expensive (see e.g. http://www.enablemart.com/Voice-Recognition). Besides, if the transcribed discourse is multilingual, additional technical problems have to be solved. So, in most cases the transcribing is to be performed manually for the time being.

English language resources dominate in **spoken corpora**, which is quite predictable. It would be enough to mention Cambridge International Corpus (CANCODE, http://www.cambridge.org/elt/corpus/cancode.htm), Diachronic Corpus of Present-Day Spoken English (DCPSE, http://www.ucl.ac.uk/english-usage/projects/dcpse/index.htm), Michigan Corpus of Academic Spoken English (MICASE, http://quod.lib.umich.edu/m/micase/). A considerable list of spoken corpora can be found at http://corpus-linguistics.de/html/corp/corp_spoken.html. Compiling of non-English spoken corpora is lagging behind. The research presented in this paper deals with two languages, Finnish and Russian, which are no exception. The Russian National Corpus includes a spoken subcorpus of about 6 million running words (www.ruscorpora.ru, [3]). Transcripts of Finnish speech are available from the Finnish Broadcast Corpus (Finnish Bank of Language, http://www.csc.fi/english/research/software/fbc). Most of the other existing collections of transcripts of prepared and spontaneous speech are of modest size and with only basic search interface or no search interface at all.

Not surprisingly, the tools and methodology used in spoken corpus research are developed along the same lines as the tools for processing written language. The transcripts are regarded as a sort of written texts. Many of the spoken corpora do not even use any transcribing conventions (e.g. MICASE).

The research of interpreting is a quite important part of translation studies. However, **interpreting corpora** are still quite a new kind of language resources and thus far not much quantitative data is available. We would like to mention the European Parliament Interpreting Corpus (EPIC) as one of the

[1] Of course, corpus tagging and annotation is still a problem, and corpus research is still limited by the lack of annotated corpora.

very few examples. The corpus consists of speeches at the European Parliament interpreted into English, Italian, and Spanish, and is arranged as a parallel corpus (http://sslmitdev-online.sslmit.unibo.it/corpora/corporaproject.php?path= E.P.I.C.). The lack of the language resources makes it difficult to obtain extensive research data, to say nothing of data processing facilities pertaining to interpreting. Consequently, there is a huge demand for more electronic data to be employed in interpreter training and interpreting research.

Besides conventional converting of written text in one language into written text in another language (translation) and converting of oral speech in one language into oral speech in another language (interpreting) there exist other types of translating. Written text may be interpreted impromptu, speech may be translated into a text by means of subtitling or speech-to-text reporting[2].

Subtitling is an important type of non-conventional translating especially in the countries where it is not common to dub movies and tv-programs. It is important to mention, that subtitling is in many ways different from conventional written translation, see e.g. [4]. With the growth of the market for audiovisual products, subtitling has become an object of research, and corpus data is needed. The Open Source Parallel Corpus (OPUS) includes a parallel corpus of subtitles (http://urd.let.rug.nl/tiedeman/OPUS/cwb/ OpenSubtitles/frames-cqp.html) [5].

In this paper, another important kind of resources is introduced. The data is arranged as parallel corpus with speech and subtitles aligned. The Subtitling corpus we are designing presents a new type of language resource with both the speech transcripts and the subtitles included. When investigating subtitled material it is important to have access not only to a film script but to the entire audiovisual message. With that purpose in mind, this corpus should consist of an exact transcript of the film dialogue as well as the relevant information on its other auditive and visual elements. This data would be aligned with the subtitles. We know nothing about existing corpora of this kind.

II. RESEARCH METHODS AND MATERIAL

A. Research Methods

The research of the data supplied in the interpreting corpora shall not be confined to examining corresponding passages in the original speech and in the speech of the interpreter. The holistic approach taken would study the communication between the participants and the interpreter, the message transmission via interpreting, the communicative failures during interpreting, the extralinguistic activities of the communicants, etc. The audiovisual translation analysis would also take into account the visual channel and the pressure on the recipient, who has to read the subtitles at the same time as s/he watches the movie. Apart from methods in functional theories of translation, some directions of established

[2] A new form of communication used for communication between deafened and hearing people.

linguistic theory will also be suitable in analyzing and interpreting the research results: discourse and conversation analysis, linguistic pragmatics, theory of speech acts, etc.

B. Research Material

A number of Finnish-Russian electronic corpora: the Corpus of court interpreting (CIC), the Corpus of learners' interpreting (CLI), and the Corpus of film transcripts and subtitles (FiTS), are currently being collected and placed on the web site of the project.

The structure of the database is established and a pilot version of the search engine for the spoken corpora has been developed. The data is currently stored on the server of the Russian Section of the Department of Translation Studies (https://mustikka.uta.fi/spoken/, access restricted to the members of the research team).

III. COMPOSITION OF THE CORPORA

In institutional interpreting contexts, part of communication often takes place in one language without the help of the interpreter, who takes part in discussion when needed. Even when the interpreter does take part in the communication, the process is often not as smooth, as it might be expected. The speakers often interrupt each other, and the interpreter works under constant pressure. The interpreting discourse is thus a sophisticated mixture of verbal and non-verbal communication, part of which is mediated by the interpreter (see e.g. [6]–[8]).

The same features can be found in a film with subtitles. It is a very complicated stream of information: visual images, sounds, verbal and non-verbal communication of the characters, speech of the narrator, text as part of original film, and subtitles (see [9]). Subtitling is a very specific kind of activity, and the subtitler must be aware both of communication problems and technical issues (see [10] and [11]).

In many respects these two kinds of data – interpreting discourse and a film with subtitles – can be reproduced in corpus databases of the same structure. Such a corpus can be arranged as a hybrid of a bilingual corpus and a parallel one. Thus, an interpreting corpus would be a collection of transcripts of speech in two (or even more) languages, and some of the transcripts would be aligned [12]. A corpus of film transcripts and subtitles would be a synthesis of a spoken corpus (transcripts), a text corpus (subtitles), and a parallel corpus (aligned transcripts and subtitles).

Audio and visual components would in many cases be extremely useful additions to the corpus data. Unfortunately, it is not always possible to include them due to problems of ethical, copyright, and technical nature. However, remarks and comments are seriously considered as a part of corpus structure. All above mentioned issues make the architecture of the corpus quite sophisticated and the mark-up vitally important.

The transcripts are annotated using xml markup. The transcription is broad; however, speech is not smoothed up to

written language as it happens in many projects, which do not directly contribute to linguistic research. We mark pauses and their lengths as well as some prosodic features (logical accent, rising/falling pitch, etc.). No punctuation marks are used in the transcripts but question and exclamation marks, which make reading easier. The features relevant from the point of view of translation process are also subject to markup, these are deletions, additions, changes, etc.
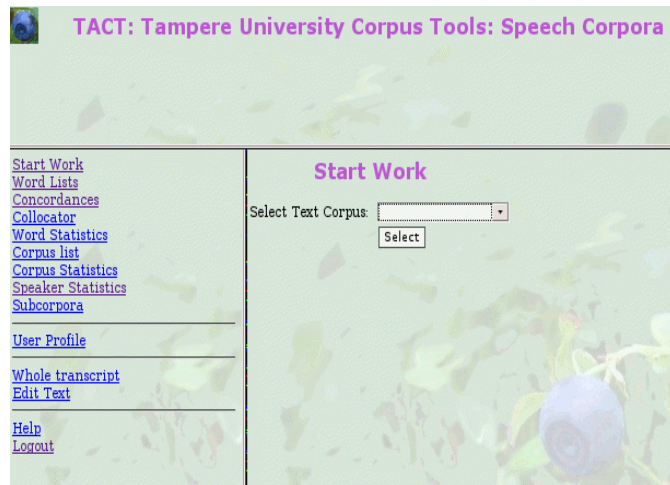

Fig. 1. TACT: User interface.

Nonetheless, xml document is not the final representation of the corpus, which is stored in a database format. The reason why transcripts are not fed into the database directly is the relative ease of markup in xml, which can be done in any word processor. It is also quite a simple task to check the consistency of the markup. So, the data extracted from xml files are uploaded to Postgresql databases (http://www.postgresql.org). The database handles many different routines like data maintenance, search, corpus users, sessions, etc. The most important for the search engine database tables are the following:

− Transcripts. Each running word, pause, tag is stored in a separate record. This makes it possible to build concordances, word lists, calculate statistics using SQL queries.
− Phrases. The start and end of each phrase is marked in Transcripts table with special tags and all the data on the phrase (speaker, timing, duration, etc.) are stored in a separate table.
− Lemmas. The lemmas of the word tokens are stored in separate tables linked to the Transcripts table. This makes the Transcripts table more transparent, saves space on disk, and simplifies generating of lemmatized lists. Lemmatization is performed after tokenization with the help of external software. English and Finnish texts are lemmatized with Connexor software (http://www.connexor.eu/technology/machinese/ machinesephrasetagger/), German with Morphy (http://www.wolfganglezius.de/ doku.php?id=cl:morphy, [13]), Russian with Rmorph

(http://www.cic.ipn.mx/~sidorov/rmorph/index.html, [14]).
− Library. The information on each item of the corpus (e.g. a film, an interview, a hearing at the Court, etc.) is stored in a separate table. The data available is text code in the corpus, title, author (for written text), date of issue, as well as text statistics (number of characters, number of running words, etc.).

IV. CORPUS TOOLS

The maintenance of the corpus database (tokenization, lemmatization, updating statistics, etc.) is performed by running php-scripts in terminal window.

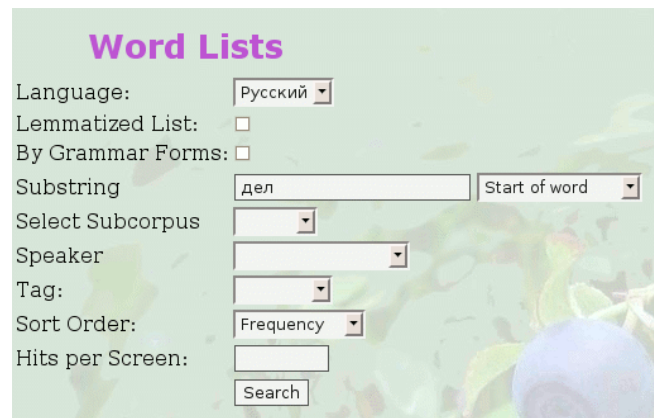The most important and frequently used search routines are

Fig. 2. Word Lists. Dialog.

included into the TACT web interface (Tampere University Corpus Tools, developed by Mikhail Mikhailov). Not surprisingly, a written-language bias in the tools and methodology of spoken corpus research is quite obvious. We mean that same tools are used for processing spoken corpora as for the written ones. Some of the spoken corpora do not even use transcribing conventions (e.g. MICASE). The TACT package also includes routines, which can be used for processing of written texts as well. However, certain functions were developed specially for spoken corpora. The software package is being constantly modified, and new functions are added to meet the requirements of the research team. Most of the functions work both with the whole corpus and with subcorpora (i.e. groups of texts defined by the user). The following research tools are currently available:

The following tools are currently available:
− Word lists;
− Concordances;
− Collocation lists (though not very helpful due to the modest size of the corpus);
− Corpus statistics;
− Speaker statistics
− etc. (see Fig. 1).

Some of the tools are more relevant for processing written texts; some have been substantially revised for the purpose of studying speech transcripts.

A. Word Lists

This tool is more flexible than standard applications for building word lists.

Word Lists Search Results

New Query

Save to file

Word	Abs. frequency	Rel. frequency
дело	9	0.38
делал	7	0.30
дела	6	0.26
делали	4	0.17
делать	3	0.13
деле	3	0.13
дел	2	0.09
делах	1	0.04
делись	1	0.04
деликатесный	1	0.04
делу	1	0.04
деловых	1	0.04

Fig. 3. Word Lists. Search Result.

It is possible to generate frequency lists of word tokens, running words, grammar tags, or even frequency lists of tokens marked by certain tags. The utility generates frequency lists for the whole corpus or for a subcorpus. Sometimes it might be quite helpful to obtain a frequency list for a certain speaker. It is no need to waste time on generating the whole list if the researcher is interested only in most frequent words, or in the words following certain pattern. On Fig. 2 the user is requesting a Russian unlemmatized frequency list of words starting with string *del* ordered according to frequency.

The resulting frequency list is displayed on Fig. 3; in addition to the absolute frequencies the relative frequency per 1000 words is calculated and shown as well.

B. Concordances

It is much more difficult to present a readable concordance derived from a speech transcript than from a written text. Moreover, the interpreter's speech has to be detached from the source speech.

The solution we suggest is to use two-column presentation with source speech in the left column and interpreting in the right one (see Fig. 4). The rise and fall of the tone, emphasis and other prosodic features are also visualized. The problem of presenting speech overlapping remains unsolved; overlaps are currently marked with brackets, which is not very user-friendly.

C. Speaker Statistics

The most interesting research tool of the TACT application is the utility presenting speaker statistics. It calculates speaker's speech tempo, number of pauses, length of pauses and other parameters. For the interpreter the script calculates statistics separately for all languages he/she speaks during the hearings.

This tool is significant for studying interpreting, whereas it is less relevant with subtitling, although it might be of use in linguistic research of film transcripts.

V. CURRENT STATE OF THE PROJECT

The corpora are currently being collected at the School of Modern Languages and Translation Studies of the University of Tampere as graduate and post-graduate research.
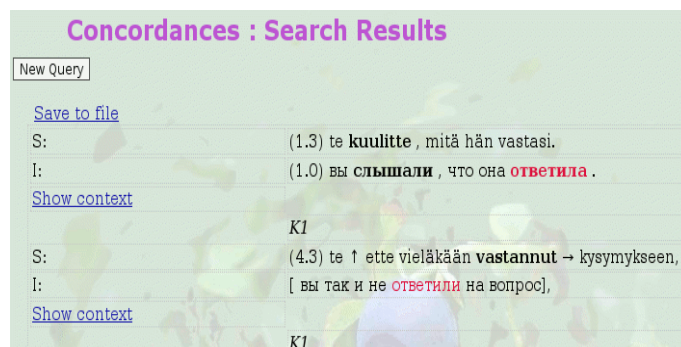
Court Interpreting Corpus. Currently nine hearings (about 48,989 running words) have been transcribed, tagged, and placed on the server.

Corpus of Film Transcripts and Subtitles: Three films (*Brat / The Brother, Kukushka / The Cuckoo*, and *Osobennosti natsional'noj ohoty / Peculiarities of the national hunting*) have been transcribed and aligned with the Finnish subtitles.

Corpus of Learners' Interpreting: Two talks with consecutive interpreting by three students of the Russian Translation Studies have been recorded transcribed and uploaded to the corpus database. Although the corpus is quite small, 12,531 running words, it is richly annotated with additions, deletions, and changes tagged.

Although the amount of material is modest, it is unique in many respects, and some interesting results have already been obtained. However limited the available data is, it enables some implementation of quantitative methods in studying interpreting and subtitling.

The project also sets new challenges in developing an efficient, robust and flexible search engine for processing interpreting and subtitling corpora, i.e. electronic corpora with transcripts of discourse with consecutive and/or simultaneous interpreting or subtitles.

Concordances : Search Results

New Query

Save to file

S:	(1.3) te **kuulitte** , mitä hän vastasi.
I:	(1.0) вы **слышали** , что она ответила .
Show context	
	K1
S:	(4.3) te ↑ ette vieläkään **vastannut** → kysymykseen,
I:	[вы так и не ответили на вопрос],
Show context	
	K1

Fig. 4. Concordance.

Speaker Statistics

Save to file

Speaker	Number of Words	Number of Phrases	Pauses: number / duration	Time	Average speech tempo
K1					
EAo, fi	34	4	21 / 35.7 s.	00:00:57	0.60/1.60
EV, fi	495	32	95 / 39.8 s.	00:03:26	2.40/2.98
I, fi	1312	119	459 / 159.04 s.	00:12:50	1.70/2.15
I, ru	1351	130	354 / 109.3 s.	00:16:19	1.38/1.55
S, fi	775	91	265 / 285.5 s.	00:11:22	1.14/1.95
T, fi	157	23	39 / 40.9 s.	00:01:42	1.54/2.57
V, ru	1591	128	315 / 204.8 s.	00:15:32	1.71/2.19
XX, fi	25	6	4 / 3.6 s.	00:00:13	1.92/2.66
K2					
Com, fi	34	1	4 / 1 s.	00:00:10	3.40/3.78
EAo, fi	36	2	14 / 6.9 s.	00:00:19	1.89/2.98
EV, fi	355	28	80 / 42.1 s.	00:02:53	2.05/2.71
I, fi	638	26	259 / 87.4 s.	00:06:49	1.56/1.98
I, ru	99	12	33 / 14.2 s.	00:00:56	1.77/2.37
S, fi	46	6	27 / 19.6 s.	00:00:38	1.21/2.50
T, fi	51	7	11 / 9.7 s.	00:00:29	1.76/2.64
To, fi	183	31	35 / 16.6 s.	00:01:29	2.06/2.53
To, ru	847	33	108 / 47.6 s.	00:04:55	2.87/3.42
XX, fi	4	3	3 / 4.2 s.	00:00:06	0.67/2.22

Fig. 5. Speaker statistics.

Subtitling is a major means of inter-cultural communication and an extremely widely read text type in 'subtitling countries' such as Finland. There is a great need for systematic data to help improve subtitle quality and understand the subtitling process and audience expectations. The parallel corpus of transcripts and subtitles, which combines spoken and written data, is an entirely new type of language resource promising an important step forward in subtitling research.

We believe that the corpora can be used both directly and indirectly in Interpreting and Translation Studies: in training of interpreters and subtitlers, and in theoretical descriptions of the structure of multilingual and multimediated discourse.

REFERENCES

[1] J. Pomikálek, P. Rychlý and A. Kilgarriff, *"Scaling to Billion-plus Word Corpora,"* in *Advances in Computational Linguistics. Special Issue of Research in Computing Science,* Vol 41, Mexico City. 2009. Available: http://www.kilgarriff.co.uk/Publications/2009-PomikalekRychlyKilg-MexJournal-ScalingUp.pdf

[2] E. G. Devine, S. A. Gaehde, and A. C. Curtis, "Technology Evaluation: Comparative Evaluation of Three Continuous Speech Recognition Software" in *Packages in the Generation of Medical Reports JAMIA* 2000, pp. 462-468.

[3] E. Grišina, "Ustnaja reč v Nacional'nom korpuse russkogo jazyka," *Nacional'nyj korpus russkogo jazyka:* 2003—2005. M.: Indrik, 2005.

[4] Y. Gambier, "Challenges in research on audiovisual translation," in *Translation research projects*, Tarragona, 2009, pp. 17—27.

[5] J. Tiedemann, "Improved Sentence Alignment for Movie Subtitles," in *Proceedings of RANLP '07*, Borovets, Bulgaria, 2007. http://urd.let.rug.nl/tiedeman/OPUS/.

[6] R. González, V. F. Vásquez, H. Mikkelson, *Fundamentals of Court Interpretation. Theory, Policy, and Practice,* Durham, North Carolina: Carolina Academic Press, 1991.

[7] S. Hale, *The Discourse of Court Interpreting. Discourse practices of the law, the witness and the interpreter,* Amsterdam Philadelphia: John Benjamins, 2004.

[8] T. R. Välikoski, *The Criminal Trial as a Speech Communication Situation,* Tampere: Tampere University Press, 2004

[9] A. Rosa, "Features of Oral and Written Communication in Subtitling," *Multimedia Translation,* Y. Gambier and H. Gottlieb (eds.), John Benjamins, Amsterdam/Philadelphia, 2001.

[10] J. Heulwen, "Quality Control of Subtitles: Review or Preview," *Multimedia Translation*. Y. Gambier and H. Gottlieb (eds.), John Benjamins, Amsterdam/Philadelphia, 2001.

[11] J. Pedersen. "Scandinavian Subtitles: A comparative study of subtitling norms in Sweden and Denmark with focus on extralinguistic cultural references," Ph.D. dissertation. Stockholm: University of Stockholm. 2007.

[12] M. Mikhailov and N. Isolahti, "Korpus ustnyx perevodov kak novyj tip korpusa tekstov (The corpus of interpreting as a new type of text corpora, in Russian)," in *Dialog-2008 International Conference*, June 4th–8th, Moscow, 2008, http://www.dialog-21.ru/dialog2008/materials/html/58.htm.

[13] W Lezius, "Morphy - German Morphology, Part-of-Speech Tagging and Applications," in *Proceedings of the 9th EURALEX International Congress* Stuttgart, Germany, 2000, pp. 619-623. Available: http://www.wolfganglezius.de/lib/exe/fetch.php?media=cl:euralex2000.pdf

[14] A. Gelbukh and G. Sidorov, "Approach to construction of automatic morphological analysis systems for inflective languages with little effort. In: Computational Linguistics and Intelligent Text," *Lecture Notes in Computer Science*, N 2588, Springer-Verlag, 2003, pp. 215–220.