# Computers in the Yugoslav Serbo-Croat/English Contrastive Analysis Project

Željko Bujas, Ph.D.
Assistant Professor
Department of English
Zagreb University, Zagreb, Yugoslavia

0.1. As far as the present writer is aware, the Yugoslav Serbo-Croat/English Contrastive Analysis Project[1] is the first contrastive analysis effort to use a large corpus of parallel texts. The corpus is made up of the Brown Corpus (reduced by 50%) with its Serbo-Croat translation, and a smaller Control Corpus (Serbo-Croat originals and English translation). A total, thus, of twice 500,000 words plus twice 150,000 words, or a grand total of some 1,300,000 words of running text.

0.2. The Project, let us make it clear, is not exclusively based on this corpus. Compilation and confrontation of grammatical statements by various authors, plus plain old intuition, figure prominently in the methodology. The insistence on a large corpus, however, is due to the conviction, prevailing among the Project workers, that only an extensive investigation of correspondences (original-language elements and their translations) can adequately reveal the less predictable patterns which tend to have a considerable contrastive analysis potential.

0.21. The most productive method of obtaining correspondences from our corpus is to concordance separately its Serbo-Croat and English parts, then to merge the resulting KWIC concordances into a contrastive KWIC concordance (with English keywords and alternating English and Serbo-Croat lines). For the more promising patterns, the merging procedure will be used twice, with both English and Serbo-Croat keywords.

0.22. In view of the size of the corpus, and the extensive concordancing required as a major procedure in the Project, the need for computer processing is obvious. It requires no undue strain on imagination to realize the soul-numbing effect of sheer physical handling of this mass of text if written out on slips.

Even in its most efficient and flexible form of a manual concordance (a sentence-slip file with keywords underlined monolingually), without which no manual pairing of correspondences is possible, the manual handling of this 1,300,000-word corpus calls for a staggering amount of time and effort to prepare. According to our careful estimate, a total of 7,100 man-hours is required to make such a concordance (without the 1,900 hours of translation from English to Serbo-Croat, and vice versa).

0.23. The slip file thus obtained would, however, secure only a one-way approach: either from English or Serbo-Croat. A slip-file allowing a two-way approach would require an additional effort of at least 4,500 man-hours.

0.24. Finally, even these two manual concordances would still leave unfilled the need for reverse concordancing, so important for morphosyntactic research. To meet this need, two additional (though less ample) slip files would have to be established.

1.0. In view of all this, the Yugoslav Serbo-Croat/English Contrastive Analysis Project has from the outset linked the planning of its work to the services of a local computer, the City of Zagreb IBM 360/30 machine[2]

1.1. Stage 1 of computer processing. The tape with the full text of the Brown Corpus (purchased from Brown University, Providence, R.I., U.S.A.), which had been prepared on an IBM 7090 machine, had first to be converted from the density of 800 BPI to 1,600 BPI, required by the Zagreb computer.

1.11. After this, a printout of the entire text was obtained on the Zagreb machine. The printing took about eight hours, with a special program[3] restructuring the original format of the Brown Corpus text. This program left out the location-marker column on the right-hand margin of printout[4], and added a sequence of sentence numbers (from 00001 to 52533) on the left.

1.12. The full text of the Brown Corpus was now reduced by 50%, retaining, however, as closely as possible, the same proportions of the 15 genres (styles) contained in the Corpus.

1.13. Printouts of the samples retained in this reduced version were then sent out to reliable translators, selected to be representative of the three major regional variants of Serbo-Croat (western, central and eastern). Their instructions were to translate at normal speed, and as carefully as when they do any other paid translation work. The only limitation imposed upon them was to observe the sentence limit in the original (English or, in the used for the preparation on the IBM 360/30 of a full forward KWIC concordance of the Serbo-Croat Corpus.

1.8. Stage 8. Using the same tape, we now plan to produce a reverse KWIC concordance of the Serbo-Croat text. This concordance will be selective in the same sense that the English reverse concordance was (cf. Stage 4).

1.9. Stage 9. With the normal and reverse KWIC concordances of both the English and Serbo-Croat corpora now obtained[8], we can move on to the final stage(s) of central importance to the Project, i.e. the merging of these monolingual concordances to get contrastive concordances (cf. 0.21.). We have planned four such concordances, and have attempted to illustrate them here by short simulated samples. As at the time of writing this no concordances of the Brown Corpus text (either original or translation) were available, the text used for these samples is the Serbo-Croat original and its translation into English of the novel Povratak Filipa Latinovicza (The Return of Philip Latinovicz) by the contemporary Croat writer Miroslav Krleža.

### 1.91. Forward contrastive concordance[9] (English to Serbo-Croat)

```
0003 ND THE DOOR LOCKED, AND HIMSELF SHUT OUT IN THE STREET, AND EVER SINCE THEN
0003 ASZAO ZAKLJUCYANA VRATA I OSTAO NA ULICI, TE OTADA ZZIVI NA ULICI VECZ MNOG

0012 E TONGUE OF THE COKE FLAME FLICKERED OUT FROM UNDER THE PAINTED IRON STOVE/
0012 STI JEZICYAC KOKSOVOG PLAMENA POD STALKOM NASLIKANE GVOZDENE PATENT-PECZI/2

0040 RIANGS STATUE AND THEY NEVER GOT HIM OUT, AND THE WATER ABOVE HIM WAS STAIN
0040 URIJANA I DA GA VISZE NIKADA NISU IZVUKLI, NEGO SE JE SAMO VODA ZAKRVARILA

0044 G WHEN HIS OWN MOTHER HAD TURNED HIM OUT INTO THE STREET IN MORAL INDIGNATI
0044  JUTRA, KADA GA JE RODZENA MAJKA IZBACILA NA ULICU S MORALNIM ZGRAZZANJEM,

0088 HE DISTANCE, EVERYTHING WAS SWELLING OUT IN THE SILENT INSTRUMENTATION OF T
0088 U DALJINAMA, SVE JE RASLO KAO TIHA INSTRUMENTACIJA MODROG JUTARNJEG BUDZENJ

0190 WITH ITS ROLLS OF BREAD, -- ALL GAVE OUT THE ACRID AND PUNGENTLY ACRID SMEL
0190 EMLJAMA, KAO KOPRENA/1/ IZ SVEGA STRUJI OSZTAR I OSJETLJIVO VLAZZAN VONJ DU

0196 ARLIER, THE STUFFING HAD BEEN COMING OUT, A MASS OF BANDS, CURLY FEATHERS A
0196 ET GODINA, PROVIRIVALA UTROBA, ISPUNJENA GURTAMA, PERASTIM KOLUTIMA I CYUPE

0210                     FIRE BLAZES OUT OF THE IRON THROATS AND THERE IS A
0210                SUKLJA OGANJ IZ ZZELJEZNIH ZZDRIJELA I MIRISZE BARUT/1/ JE

0240           A WAX CANDLE WAS BURNING OUT ON A MARBLE SQUARE OF THE CHURCH F
0240           DOGORIJEVALA JE NA MRAMORNOJ CYETVORINI CRKVENOG PODA JEDNA VOS

0271 RY HUMAN EYE, LIKE AN ANIMAL PEEPING OUT OF A CAGE/2/ HUMAN GESTURES ARE LI
0271 UDSKOM OKU IMA TUGE, KAKVOM DOGADZAJE PROMATRAJU ZZIVOTINJE IZ KAVEZA/2/ KR
```

Control Corpus, in Serbo-Croat). They were not to split the English sentence into two or more Serbo-Croat sentences, nor were they allowed to combine two or more English sentences into one Serbo-Croat sentence.

The reason for this was the need to secure a mechanical pairing of the English (or Serbo-Croat) keyword, marked by its sentence number, with the same-numbered, parallel, Serbo-Croat (or English) sentence in the two-language concordancing planned for the later Project stages.

1.2. Stage 2. A new magnetic tape will be prepared of the reduced Brown Corpus text, and with the sentence sequence numbers interpolated. This version will be used for all subsequent concordancing.

1.3. Stage 3. Using this magnetic tape, the IBM 360/30 will now prepare a full forward KWIC concordance of the reduced Brown Corpus text.[5]

1.4. Stage 4. Now (while the reduced Brown Corpus is still being translated) we shall use the same tape to obtain a reverse KWIC concordance of the same text. Since all "function words" - such as *of*, *had*, *most*, *those*, *did*, etc. - were already isolated in the previous stage (in the forward concordance)[6] this will further reduce the mass of text to be concordanced by one-half.[7]

1.5. Stage 5. The Serbo-Croat translation of the reduced Brown Corpus, by now in an advanced stage, will be copied out on a Flexowriter in batches (as translators send in their typescripts), resulting in a paper tape.

1.51. The same procedure can, at this stage, be applied to the 300,000 words of the Control Corpus. No time for translation has to be set apart here, since only already published English translations of Serbo-Croat originals are to be used.

1.6. Stage 6. Although the Serbo-Croat paper tapes obtained in the preceding stage are immediately computer-processable, we shall convert them to a magnetic tape, because this medium secures an incomparably speedier processing on the computer.

1.61. We hope that stages 2 to 6 will not take more than twenty weeks (if enough personnel can be hired simultaneously).

1.7. Stage 7. The Serbo-Croat magnetic tape will now be

## 1.92. Forward contrastive concordance (Serbo-Croat to English)

```
2205 BUKVOM, GDJE SU SE BILI SKLONILI ONE BURNE NOCZI, POSLIJE ROKOVOG PRUSZTENJ
2205 LD OAK-TREE WHERE THEY HAD FOUND SHELTER THAT STORMY NIGHT ON THEIR WAY BAC

2144 AJJ LIJECYNICI U SVOJIM TAJANSTVENIM BURNUSIMA /SZTO IZGLEDAJU KAO STAROMOD
2144 MOVED PHYSICIANS IN THEIR MYSTERIOUS BURNOUSES LIKE OLD-FASHIONED NIGHTSHIR

0216 ISERA, NAKOSTRIJESZENA LAVLJA GRIVA, BURSKE BATERIJE PRED LADYSMITHOM, MARS
0216 RL-DIVERS, THE LION&S BRISTLING MANE, THE BOER BATTERIES AT LADYSMITH, THE

2144  MIRISZU JE, IMA LI U NJOJ KARAMELA, BUSZE MU PO ZUBIMA, MJERE MU TLAK KRVI
2144 S AND SMELLING IT TO FIND OUT WHETHER THERE WAS ANY SUGAR IN IT, DRILLING H

0984                   &&PROSIM VAS, JAGO, BUTE SPAMETNI/4/
0984                                    &&PLEASE, YAGA, BE SENSIBLE/4/

1546 SMATO TALASANJE GUZOVA I LISNJACYA I BUTINA, DEBELIH MASNIH ZZENSKIH NOGU,
1546  HAIRY BUTTOCKS AND CALVES AND THIGHS, FAT WOMEN&S LEGS, ANKLES, JOINTS, SK

0210 AKAVA STEGNA KONJSKA, KRVAVE RANJENE BUTINE, UZNEMIRENE CRNE REPINE, RASKRV
0210 LANKS, BLOOD-STAINED AND WOUNDED, THEIR LONG BLACK LASHING TAILS, THEIR ELL

0473 EKANE POJASE MESA OKO KUKOVA I IZNAD BUTOVA U CJELINI POTEZA, A OVAJ TU &&B
0473 SOFT ROLLS OF FLESH ROUND THE HIPS AND ABOVE THE THIGHS, WHILE THIS FELLOW

1314 AVODLAKAVOJ OBLINI KONJSKIH STEGNA I BUTOVA, TO JE JEDINI VELIKI DOZZIVLJAJ
1314 INING HAIRY FLANKS AND HINDQUARTERS, HAD BEEN THE ONLY GREAT EXPERIENCE OF

0248 LAVE, ZZALOSNE PTICYJE OCYI, KRAVLJE BUTOVE, KONJSKA STEGNA, A SINOCZ JUSZ
0248 SE LEGS, WRETCHED BIRDS& WINGS, COWS& BUTTOCKS, HORSES& HAUNCHES, WHILE ONE
```

## 1.93. Reverse contrastive concordance (English to Serbo-Croat)

```
0002 WAS ALL STILL FAMILIAR TO HIM/1/ THE ROTTING, SLIMY ROOFS, THE ROUND BALL O
0002 NAO JE JOSZ UVIJEK SVE KAKO DOLAZI/1/ I TRULI SLINAVI KROVOVI I JABUKA FRAT

0003 NTY-THREE YEARS HAD PASSED SINCE THE MORNING WHEN HE HAD SLUNK UP TO THAT D
0003 DESET I TRI GODINE SU PROSZLE OD ONOG JUTRA, KADA SE DOVUKAO POD OVA VRATA

0003 EET, AND EVER SINCE THEN HE HAD BEEN LIVING IN THE STREET, AND NOTHING HAD
0003 LICI, TE OTADA ZZIVI NA ULICI VECZ MNOGO GODINA, A NISZTA SE NIJE PROMIJENI

0003 E HAD BEEN LIVING IN THE STREET, AND NOTHING HAD REALLY CHANGED.
0003 VI NA ULICI VECZ MNOGO GODINA, A NISZTA SE NIJE PROMIJENILO UGLAVNOM.

0004 DLY LOCKED DOOR AND, JUST AS ON THAT MORNING, HE COULD FEEL THE COLD, IRON
0004 M ZAKLJUCYANIM VRATIMA, I KAO I ONOG JUTRA IMAO JE OSJECZAJ HLADNOG, GVOZDE

0004 AS HE PUSHED IT, HOW THE LEAVES WERE QUIVERING IN THE UPPER BRANCHES OF THE
0004 NJEGOVOM RUKOM I ZNAO JE, KAKO SE LISZCZE MICYE U KRUSZNJAMA KESTENOVA I CY

0004 AS IF IN A DREAM -- AS ON THAT OTHER MORNING -- /1/ HE WAS ALL DIRTY, TIRED
0004 ILO MU JE /ONOG JUTRA/ KAO DA SANJA/1/ BIO JE SAV CYADZAV, UMORAN, NEISPAVA

0004 RED, IN NEED OF SLEEP, HE COULD FEEL SOMETHING CRAWLING INSIDE HIS COLLAR -
0004 RAN, NEISPAVAN, OSJECZAJUCZI KAKO MU NESZTO PLAZI OKO OKOVRATNIKA/1/ PO SVO

0004 ED OF SLEEP, HE COULD FEEL SOMETHING CRAWLING INSIDE HIS COLLAR -- A BED-BU
0004 N, OSJECZAJUCZI KAKO MU NESZTO PLAZI OKO OKOVRATNIKA/1/ PO SVOJ PRILICI STJ

0005 RO, LAST DRUNKEN NIGHT, AND THE GREY MORNING.
0005 PIJANE, POSLJEDNJE, TRECZE NOCZI I ONOG SIVOG JUTRA -- DOK ZZIVI.
```

## .94. Reverse contrastive concordance
(Serbo-Croat to English)

```
0002 RVOREDA, MEDUZINA GLAVA OD SADRE NAD TESZKIM, OKOVANIM HRASTOVIM VRATIMA I
0002 LASTER HEAD OF MEDUSA SURMOUNTING THE HEAVY, IRON-BOUND OAK DOOR WITH ITS C

0002 MEDUZINA GLAVA OD SADRE NAD TESZKIM, OKOVANIM HRASTOVIM VRATIMA I HLADNA KV
0002  OF MEDUSA SURMOUNTING THE HEAVY, IRON-BOUND OAK DOOR WITH ITS COLD LATCH.

0002 GLAVA OD SADRE NAD TESZKIM, OKOVANIM HRASTOVIM VRATIMA I HLADNA KVAKA.
0002 USA SURMOUNTING THE HEAVY, IRON-BOUND OAK DOOR WITH ITS COLD LATCH.

0004                           ZASTAO JE PRED STRANIM ZAKLJUCYANIM VRATIMA, I KAO I
0004          HE STOPPED IN FRONT OF THE UNFRIENDLY LOCKED DOOR AND, JUST AS ON T

0004                   ZASTAO JE PRED STRANIM ZAKLJUCYANIM VRATIMA, I KAO I ONOG JUT
0004  HE STOPPED IN FRONT OF THE UNFRIENDLY LOCKED DOOR AND, JUST AS ON THAT MOR

0006  GDJE SE JE KAO MALI DECYKO IGRAO SA SVOJIM BIJELIM JANJCEM, STAJALO JE GRA
0006 ERE AS A BOY HE HAD PLAYED WITH HIS WHITE LAMB, THERE WAS A BUILDING-SITE W

0006 E JE KAO MALI DECYKO IGRAO SA SVOJIM BIJELIM JANJCEM, STAJALO JE GRADILISZT
0006 S A BOY HE HAD PLAYED WITH HIS WHITE LAMB, THERE WAS A BUILDING-SITE WALLED

0006  JE GRADILISZTE OBZIDANO KAO CYOVJEK VISOKIM ZIDOM I NA TOM VISOKOM ZIDU BI
0006 -SITE WALLED IN LIKE A MAN BEHIND A HIGH WALL, AND ON THIS HIGH WALL THERE

0009                         DUGO JE STAJAO POD VITKIM ZZENSKIM STEZNICIMA, A PRSTI SU
0009 RE FOR A LONG TIME UNDER THE SLIM CORSETS, AND HIS FINGERS WERE ALL DIRTY W

0009                 DUGO JE STAJAO POD VITKIM ZZENSKIM STEZNICIMA, A PRSTI SU MU BIL
0009 A LONG TIME UNDER THE SLIM CORSETS, AND HIS FINGERS WERE ALL DIRTY WITH DUS
```

.10. The reason why these four concordances have been resented under one processing stage (9) is that, first, e are not sure whether we can afford the computer for ach of them, and, second, we do not, at this point, know ow selective each of them is going to be. A considerable eduction of the text to be concordanced can be achieved n reverse concordancing if we restrict ourselves only to ords, ending in a characteristic morpheme with clearly oreseeable contrastive analysis potential (such as -ed, ly, -est, -ing, -ness, -less, etc. in English, and -ao, vši, -en, -šcu, -ost, -šte, etc. in Serbo-Croat).

.11. It may be pointed out here that, irrespective of ow restrictive the selection of keywords for concordancng may have to be, no concessions should be made in the rinciple of bilingual approach. Only if, in our investiation of the contrastive potential of individual eleents, we strictly observe the approach from both the nglish and the Serbo-Croat texts, can we be certain that e shall have covered all possible contrastive description atterns based on correspondences in both corpora.

.0. Once contrastive concordancing has been completed we shall still be facing some practical technical problems.

2.1. Project analysts, for instance, will often have to be provided with slips instead of computer printout sheets. Only if the material being analyzed is in the form of slips will they be able to classify and reclassify the key elements swiftly and flexibly (by putting together, breaking up and re-establishing batches of slips).

2.11. Cutting up the concordance printouts to get the slips is not very practical in view of the varying size of contrasted pairs of elements with their context (cf. n. 9, second half). The way around this, clearly, is to have the pairs printed out at regular intervals with sufficient blank space in between. This, however, would probably triple the amount of printout paper required. Also, this is complicated further by the need for a number of copies for each pair (slip), because of simultaneous demands that may often be made upon the same slip by several Project analysts, approaching the same element from various descriptive levels. These copies could be secured by using special, multiple-carbon printout paper, but this might prove quite expensive.

2.2. In view of all this, the Yugoslav Serbo-Croat/ English Contrastive Analysis Project has envisaged the use of a Flexowriter here as an alternative method. This machine has already provided us with the paper tape of the Serbo-Croat translation of the reduced Brown Corpus, plus the tapes of Serbo-Croat originals and English translations of the Control Corpus (cf. Stage 5). The missing paper tape of the English text of the Brown Corpus can be obtained on a magtape-to-papertape converter. Once both paper tapes are ready, running them through the Flexowriter provides us with up to 13 (some claim 20) carbons of each contrasted pair. An additional advantage of using the Flexowriter for slip duplication is in the less awkward shape of slips. Paper tapes reproduce the text in 60-character-wide lines of the original translators' typescript, as opposed to the 110 to 120-character streamers of normal computer printout (unless the concordance printout was programmed for a narrower format, requiring considerably more paper).

2.3. The resulting slip files of sentence-numbered English and Serbo-Croat texts, coupled with the Project's basic (monolingual - forward and reverse) concordances, can now be used as a replacement for contrastive concordances. It would work approximately like this: upon receiving an analyst's request for examples of all correspondences in the corpus of an element under analysis, the Project headquarters in Zagreb would look the element up in one of the basic concordances, record sentence numbers of all the occurrences, extract slips bearing these numbers from the Flexowriter-produced slip file, and forward them to the analyst for further research.

## Footnotes

1. Launched in 1968, at the Institute of Linguistics, Faculty of Arts and Letters, Zagreb University. Director: Professor Rudolf Filipović, Ph.D. (Postal address: Jugoslavenski projekt za kontrastivnu analizu srpskohrvatskog i engleskog jezika, Institut za lingvistiku, Filozofski fakultet, Djure Salaja 3, Zagreb, Yugoslavia). Project analysts, numbering 20, are on English department staffs from all parts of Yugoslavia.

2. Size of storage: 32K. Other equipment: three 2311 discs, two 2415/4 tape drives, one 2540 card reader, one 2671 paper-tape reader, one 1403/2 printer.

3. Written by Dipl.ing. Milutin Cihlar, Chief Programmer of the Zagreb system.

4. Cf. *Manual of Information* (for the Brown Corpus), Brown University, 1964, p. 7.

5. We hope to use forward and reverse concordancing programs developed by a US project for an IBM 360/30, or a similar machine.

6. In a total reverse concordance they would only appear in a different place: *of* under F, *had* under D, etc.

7. Putting the top 100 words from the Brown Corpus Rank List on the exclusion list (compared to a total of some 180 "function words", in the present author's estimate), would reduce the text by 47.4 per cent, while including only one morphologically marked word (YEARS) and two lexical words (NEW, TIME). Expanding the exclusion list to cover the top 200 words would probably not be economical (though only two additional morphologically marked words would be included: UNITED and STATES), because the computer would be slowed down, whereas the textual

mass would be reduced by only 6 more per cent (to 53.6 per cent.

8. Which may take between 40 and 60 computer hours, as opposed to an estimated 2,350 hours of manual processing (for only the English forward concordance at that).

9. In addition to being simulations, all these concordance samples are in an idealized format, with the correspondences spatially parallel to the keyword. In practice, however, it is impossible to achieve this ideal textual parallelism, because there are no other formal signals to govern it, except the sentence sequence number which can only mark the sentence as a whole.

For this reason, the actual computer concordances will, when ready, have the correspondence to the keyword printed out with the whole sentence in which it occurs, under the single line with the keyword. This will, naturally, increase the size of the concordance, but not more than about 50 per cent in our estimate. This is because only an approximate 40 per cent of all sentences in the original text of the Brown Corpus are in excess of 20 words (which can be accommodated by the average printout line). A mere 6 per cent of these sentences are longer than 40 words, requiring, consequently, more than two printout lines.