Rudolf Filipović

The Use of a Corpus in Contrastive Studies*

1.0. The first problem facing researchers engaged in a contrastive analysis project[1] is that of the method to be adopted. Immediately after that comes the closely connected question of the corpus. Obviously, the choice of method determines whether a specific corpus is needed or not.

1.1. One of the first questions that we wanted to answer before embarking upon the Serbo-Croatian—English Contrastive Project[2] was whether to base our analysis on a corpus or on native intuitions. It was clear that this question was linked with the problem of the model of description to be used in contrastive analysis.

1.2. After examining several existing contrastive studies, I found that none employed a specific and consistent method that might be regarded as *the* method of contrastive analysis.

1.3. The conclusion I drew from the literature and from our experience (based on a number of papers and theses on contrastive topics written in Zagreb seminars over several years) was that in contrastive analysis there is a strong interdependence of theory and practice, so that the best method would be one combining the theoretical and the empirical.

1.4. Our experience showed that there are areas of contrastive analysis in which no purely theoretical method would lead to a satisfactory solution.

* This paper was read at the *International Symposium of Applied Contrastive Linguistics,* at the University of Stuttgart, October 1971.

[1] R. Filipović, "Contrastive Analysis of Serbo-Croatian and English", *SRAZ,* 23, 1967, pp. 5—27.

[2] R. Filipović, "Problems of Contrastive Work", *SRAZ,* 29—32, 1970—71, pp. 19—54.

1.5. These considerations prompted us to seek a method, or a combination of methods, that would yield not only theoretical but also practical results. These practical results must be applicable in compiling and developing teaching materials and working out improved teaching methods (which is one of the basic aims of our contrastive analysis). This will only be possible if the results are set forth in a manner comprehensible to the average reader of the project's publications.

1.6. We can say that we use at the same time structuralist and transformational-generative approaches to contrastive analysis. We have concluded that a certain degree of mixture of the two is necessary. Some reports are more generative in nature than others, depending on their particular topics.

1.7. To ensure wide coverage of the linguistic phenomena involved, and to make up for the lack of linguistic theory in some areas, we adopted the translation method based on a corpus of text.

2.0. At first we laid down specific principles for the construction of our own corpus. We intended to have two corpuses (an English one translated into Serbo-Croatian and a Serbo-Croatian one translated into English) because it was clear to us from the beginning that a complete contrastive analysis based on the translation method would require two corpuses of equal size and composition. This would enable us to examine phenomena in both languages from the point of view of their translation. It soon became quite clear, however, that it would be rather difficult, if not impossible, to build a large enough corpus with the limited time and resources that we had at our disposal, and that consequently we should have to use an existing corpus and to work with only an English corpus and its Serbo-Croatian translation.

2.1. Why we have chosen the Brown Corpus of two existing corpuses (the London *Survey of English Usage* and the Brown University *Standard Sample of Present-Day Edited American English)* and how it was shortened, translated into Serbo-Croatian, grammatically coded, and finally processed by the computer, has been carefully discussed and justified in two articles.[3]

[3] R. Filipović, "The Choice of the Corpus for a Contrastive Analysis of Serbo-Croatian and English", *The Yugoslav Serbo-Croatian—English Contrastive Project, Studies* 1, Zagreb 1970, pp. 37—46.

R. Filipović, "The Yugoslav Serbo-Croatian—English Contrastive Project So Far", in R. Filipović (ed.), *Zagreb Conference on English Contrastive Projects,* Zagreb, Institute of Linguistics, 1971, pp. 31—79.

3.0. Let us now see very briefly how other contrastive projects have dealt with the question: "Should we base our analysis on a corpus or on native intuitions?".

3.1. At the *Tenth FIPLV Congress*[4] in Zagreb in 1968, in the section on *Contrastive Linguistics and Its Pedagogical Implications,* two contrastive projects were discussed. In Prof. G. Nickel's paper "Project on Applied Contrastive Linguistics"[5] PAKS was presented and in Prof. B. Carstensen's paper "Contrastive Syntax and Semantics of English and German"[6] the Mainz project was described.

3.2. The aims and tasks of PAKS were summarized as 1) "the adequate description of the German and English languages based on generative-transformational theory of grammar"; 2) " contribution to the further development of T-G theory, particularly with reference to its practical application in foreign-language teaching",[7] etc. It was evident from what we read in the paper that the method used would be T-G and that no use of a corpus was envisaged.

3.3. Prof. Carstensen's Mainz project took as its theoretical foundation the linguistic investigations carried out by Noam Chomsky. Beyond the purely scholarly interest of this research an effort would be made to emphasize the relevance of the results of contrastive analysis for teaching purposes. For the purpose of contrastive analysis this project would make reference to the great standard works on the grammar of both languages and the latest structural and transformational descriptions of their syntax.

3.3.1. Use would also be made of the great dictionaries, but besides the dictionaries, the true foundation of this research programme would be a careful examination of a maximally comprehensive corpus of the two languages under comparison. In order to accomplish this task it would be necessary to use electronic computer processing and first of all to get together a collection of textual material. It would be essential to ascertain statistically how often and with what degree of regularity certain linguistic phenomena are to be found in one particular text or in a series of different texts as the case may be.

[4] R. Filipović (ed.), *Active Methods and Modern Aids in the Teaching of Foreign Languages — Papers from the 10th F.I.P.L.V. Congress,* London, Oxford University Press, 1972, 231 pp.

[5] *Ib.*, pp. 217—226.

[6] *Ib.*, pp. 206—216.

[7] *Ib.*, pp. 225—226.

3.3.2. Prof. Carstensen also envisaged the help of informants, as "experience has already shown that some types of information can only be reliably obtained with the help of informants. Such information would be mainly on certain structures of very low frequency of occurrence, possibly stylistically determined".[8]

4.0. At the *Zagreb Conference on English Contrastive Projects*[9] (December 1970) we became acquainted with a few more contrastive projects: Polish-English,[10] Roumanian-English,[11] and Hungarian-English.[12] Each of these has developed far enough that we can refer to their points of view on the question of the method and the use of a corpus.

4.1. The members of the Polish-English project adopted the T-G model in the same year in which they began to assemble their own corpus of English and semantically corresponding Polish sentences. The sentences were taken from novels, magazines, and scientific works: 100,000 English sentences and approximately the same number of Polish sentences. The corpus is however considered only as an aid to Polish research workers.[13]

4.1.1. In 1970, Prof. Fisiak states in his report, the encoding of information concerning both English and Polish was initiated, and it should be completed by the end of 1971. This would make information concerning various aspects of the structure of English and Polish more easily accessible. The Polish-English project participants have a Polish language corpus at their disposal as well.

4.1.2. In the discussion[14] that followed Prof. Fisiak's report it was made clear, however, that "the Polish-English project considers the corpus a help in some cases, and that other cases do not require it, as the corpus is not an end in itself".[15] If it

[8] *Ib.*, p. 209.

[9] R. Filipović (ed.), *Zagreb Conference on English Contrastive Projects, 7—9 December 1970. Papers and Discussion.* Zagreb, Institute of Linguistics, 1971, 242 pp.

[10] Jacek Fisiak, "The Poznań Polish — English Contrastive Project", *ib.*, pp. 87—96.

[11] Tatiana Slama-Cazacu, "The Romanian—English Language Project", *ib.*, pp. 226—234.

[12] József Hegedüs, "Two Questions of Hungarian—English Contrastive Studies", *ib.*, pp. 101—121; László Dezsö, "Contrastive Linguistic Project on English and Hungarian in Hungary", *ib.*, pp. 124—129.

[13] J. Fisiak, *o. c.*, see note 10, p. 93.

[14] R. Filipović (ed.), *Zagreb Conference on English Contrastive Projects, 7—9 December 1970. Papers and Discussion.* Zagreb, Institute of Linguistics, 1971, pp. 97—100.

[15] *Ib.*, pp. 97—98.

furnishes only a few examples of a problem researchers can look for material outside the corpus. Another justification for this is the fact that the Polish-English project adopted the T-G model.

In further discussion it was stated that "some problems require a corpus, such as those involving norm vs. system (Prof. Coseriu). In English the topic of a sentence very often coincides with the subject, which is not so in German" (Dr. König).[16]

4.2. The Hungarian-English project in Hungary is still in its initial stage. From what we know about it we can say that it will be based on a limited corpus,[17] unless they take the Zagreb coded version of the Brown Corpus[18] and translate it into Hungarian.

4.3. In the discussion[19] that followed Dr. König's paper (in which some general questions of the method and contrastive studies were discussed) it was brought out that some contrastive projects which adopted the T-G model in the beginning have now renounced it. We heard from Dr. König that PAKS, which used to be theoretically oriented, is now much less theoretical. PAKS has also turned to a corpus in some cases.

4.3.1. Dr. König pointed out that "in investigating the problem of topicalization, in order to assess the stylistic significance of this particular phenomenon of subjectivizing certain constituents in English and in order to assess the frequency of other phenomena PAKS turned to a corpus. Or, if there is a construction in English which is less like anything in German, this construction tends to be underrepresented in the English of German speakers. In order to get these phenomena which are not a question of either-or but more-or-less one has to turn to a corpus".[20]

4.4. The English-Roumanian project plans to use a "corpus for analysis which will consist of a vocabulary of several thousand English items scientifically selected (on the basis of frequency)".[21]

4.4.1. "These lexical items", Prof Slama-Cazacu states further in her report, "will be analysed from the point of view of their

[16] *Ib.*, p. 99.

[17] Private communication of the organizers of the Project.

[18] About the numerical coding system that the Yugoslav project used in preparing the Brown Corpus see: R. Filipović, "Problems of Contrastive Work", *SRAZ*, 29—32, 1970—71, pp. 26—31.

[19] R. Filipović (ed.), *Zagreb Conference on English Contrastive projects, 7—9 December 1970. Papers and Discussion.* Zagreb, Institute of Linguistics, 1971, pp. 146—155.

[20] *Ib.*, p. 149.

[21] T. Slama-Cazacu, *o. c.*, see note 11, p. 231.

multiple meanings and the grammatical constructions in which they occur, thus arriving at the grammar that operates with this word inventory. On the basis of meaning and structure equivalencies between the languages, a similar grammar of the corresponding Roumanian lexical items will be described, thus disclosing the similarities and differences between the two languages. In describing the grammatical structure of the equivalent Roumanian words, note will be taken of their frequency, distribution, and communication value. The possible shortcomings of a corpus formed of examples drawn from dictionaries, i. e. its questionable value as a reflection of the reality of communication, will be compensated for by corroborating the results of this procedure against others directly based on the communication situation, hence on the learner".[22]

4.5. It was quite obvious at the Zagreb Conference from all the papers, and particularly from the discussion, that every project either used a corpus from the beginning or began to in the course of its work. It was very interesting to note that even the projects that were originally most theoretically oriented (like PAKS) have also turned to a corpus. In summing up the Conference,[23] I noted a recurrent theme: "The use of a corpus in contrastive analysis is not a theory and does not aim at replacing theory. The material from the corpus serves as a check on theoretically based conclusions and as a source of data in areas where the theory is inadequate".[24]

5.0. During our three years of intesive work on the Serbo-Croatian—English contrastive project I have discussed at various levels the question of the method to be used in contrastive analysis and in connection with it the use of a corpus.[25]

[22] *Ib.*

[23] R. Filipović, "Summing Up", in R. Filipović (ed.), *Zagreb Conference on English Contrastive Projects, 7—9 December 1970. Papers and Discussion.* Zagreb, Institute of Linguistics, 1971, pp. 241—242.

[24] *Ib.*, p. 241.

[25] R. Filipović, "Contrastive Analysis of Serbo-Croatian and English", *SRAZ,* 23, 1967, pp. 5—27; *idem, The Organization and Objectives of the Project,* Zagreb, Institute of Linguistics, 1968, 17 pp.; *Idem,* "Početne faze rada na projektu *Kontrastivna analiza hrvatskosrpskog i engleskog jezika. Prilozi i građa* 1, Zagreb, Institut za ligvistiku, 1969, 3—25; *Idem,* "The Choice of the Corpus for a Contrastive Analysis of Serbo-Croatian and English". *Studies* 1, Zagreb, Institute of Linguistics, 1969, pp. 37—46; *Idem,* "Contrastive Trends in Applied Linguistics", *CONTACT* 14, 1970, pp. 13—17; *Idem,* "The Yugoslav Serbo-Croatian—English Contrastive Project" (a paper read at the Second International Congress of Applied Linguistics, 8—12 September 1969 in Cambridge), in G. Nickel, ed., *Papers in Contrastive Linguistics,* Cambridge University Press 1971, pp. 107—114; *Idem,* "Problems of Contrastive Work", *SRAZ,* 29—32, 1970—71, pp. 19—54.

Two discussions (one in the United States and the other at the Zagreb Conference) made me take up the question whether or not to use a corpus in contrastive analysis.

5.1. In the first discussion the Yugoslav project was attacked for having chosen the method which required a corpus, or at least for having decided to use the chosen method and a corpus in the way I have described above. It was suggested that the final product of our work, a monograph on Serbo-Croatian—English contrastive analysis, could be written on the basis of the preliminary *Reports* discussing the topics chosen for analysis.

5.2. These *Reports* (dealing with more than fifty topics on four levels) were written by analysers on the basis of a) general works on English; b) specialized literature on each problem dealt with; c) the analyser's own knowledge and experience, and d) work with consultants. A *Report* is not, however, the final treatment of a topic. The analyser completes his report with material from the corpus by illustrating the conclusions already arrived at and by checking and supplementing results taken from the literature.

5.3. The function of the corpus in this type of work is decisive. The final results of the analysis of a topic (called *Study* in our project) depend very much on the material supplied by the corpus and only partly on the analyser's experience or the information received from the native adviser (informant).

5.4. Here are a few examples from the work of the Yugoslav Serbo-Croatian—English contrastive project to illustrate the need for a corpus in contrastive analysis.

5.4.1. The computer-processed material of the Brown Corpus was not available in the initial stage of our work, and the analyser who was discussing the English possessive adjectives[26] (*my, your, his,* etc.) could not finish his analysis without a corpus. After he had given a sketch of the topic made on the basis of the literature on the problem, he started seeking formal-semantic correspondences in Serbo-Croatian in order to analyse them and see how they differed from their English counterparts. Here he immediately felt the lack of a corpus. He had to compile a limited pilot corpus of his own to find what he called "unconditioned translation equivalence". He gave a table containing the possible groups of Serbo-Croatian equivalent va-

[26] Leonardo Spalatin, "The English Possessive Adjectives *my, your, his, her, its, our, their* and Their Serbo-Croatian Equivalents", in R. Filipović, ed., *The Yugoslav Serbo-Croatian—English Contrastive Project, Reports,* 2, Zagreb, Institute of Linguistics, 1970, pp. 94—102.

riants and the number of such groups found in his pilot corpus.[27] But all this was only provisional and statistically unreliable until he used the material from the Brown Corpus. Writing his *Study* he was able to give us all the statistics needed for a final statement on the relations between Serbo-Croatian and English possessive adjectives.[28]

5.4.2. The need for a corpus was even more evident in the analysis of the English demonstratives *this, these, that, those* and their Serbo-Croatian equivalents.[29] Here again the analyser followed the same principle. He worked on a pilot corpus of his own and the "unconditioned equivalence probability"[30] as well as the "conditioned equivalence probability"[31] shown in two tables were expressed in rather vague terms like "very little", "almost always", "according to our data" (meaning the limited pilot corpus), "quite common", etc. The figures from the pilot corpus were not statistically reliable and did not show the relations between Serbo-Croatian and English demonstratives clearly. As soon as the analyser had obtained the material from the Brown Corpus he was able to get more relevant statistics and to base his final conclusions on them.[32]

5.4.3. In the initial stage, the analyser writing a report on relative pronouns[33] used a provisional corpus of 1,000 relative clauses. The result was that some translation equivalents were not represented at all. Therefore, not only a corpus but a proper-sized corpus is required; we must be interested in numerical relations if we want to draw conclusions which can be used later in pedagogical materials.[34]

[27] *Ib.*, p. 95.

[28] This work is in progress and the analyser is writing the study on possessive adjectives using the Zagreb coded version of the Brown corpus.

[29] Leonardo Spalatin, "The English Demonstratives *this, these, that, those* and Their Serbo-Croatian Equivalents«, *ib.*, pp. 103—119.

[30] *Ib.*, pp. 106—108.

[31] *Ib.*, p. 108—115.

[32] This work is still in progress: the analyser is writing the study on demonstratives applying the Zagreb coded version of the Brown corpus.

[33] Dora Maček, "Relative Pronouns in English and Serbo-Croatian", in R. Filipović, ed., *The Yugoslav Serbo-Croatian—English Contrastive Project, Reports* 3, Zagreb, Institute of Linguistics, 1970, pp. 105—127.

[34] Cf.: Mirjana Vilke, "Teaching Problems in Presenting Relative Pronouns", in R. Filipović, ed., *The Yugoslav Serbo-Croatian—English Contrastive Project, Pedagogical Materials* 1, Zagreb, Institute of Linguistics, 1971, pp. 98—111.

Some points mentioned in the *Report*[35] which need checking on the corpus[36] are:

a) In which cases and how often relative pronouns are used, and when and why they are omitted?

b) The grammatical and semantic nature of the antecedents of relative pronouns.

c) The use of prepositions with relative pronouns and their position.

d) The grammatical function of relative pronouns.

e) The use of relative pronouns in restrictive and nonrestrictive clauses.

6.0. In English we sometimes find a splitting up of a complete constituent into an object and a subject. Sentences of the type "A tyre of the car burst" become "The car burst a tyre", or "The river burst its banks"; similarly "The car broke a wheel".[37] This is impossible in German and in Serbo-Croatian. To investigate such cases and assess the frequency, the verbs that are possible, and so on, we have to turn to a corpus or to native informants.

6.1. In another discussion[38] about non-omissible determiners in Serbo-Croatian which may be omissible in English and vice versa,[39] it was stated that it would be difficult to make a useful generalization without extensive research on a corpus.[40]

6.2. Another value of the corpus in contrastive analysis is in its educational applicability.[41] If we want to use the material

[35] Dora Maček, o. c., see note 33, p. 111 (4. *Contrastive Analysis*).

[36] The same analyser, Dora Maček, has been doing it now in writing a study on relative pronouns based on the Zagreb coded version of the Brown corpus.

[37] E. König in R. Filipović (ed.), *Zagreb Conference on English Contrastive Projects, 7—9 December 1970. Papers and Discussion.* Zagreb, Institute of Lingustics, 1971, pp. 150—151.

[38] Vladimir Ivir, "Notes on Linking Verbs and Complements in English and Serbo-Croatian", in R. Filipović, ed., *The Yugoslav Serbo-Crotian—English Contrastive Project, Reports* 5, Zagreb, Institute of Linguistics, 1971, pp. 172—183; Midhat Riđanović, "More on Linking Verb + Complement in English and Serbo-Croatian", *ib.*, pp. 184—204.

[39] (E) *He is a man of great wisdom*
(S-C) *On je čovjek velike mudrosti*
(E) *He is a man of wisdom*
(S-C) **On je čovjek mudrosti*

[40] Midhat Riđanović, *o. c.*, see note 38, p. 188 (5).

[41] The Yugoslav project and a number of other contrastive projects have pointed out the pedagogical value of contrastive analysis and its educational applicability in foreign language teaching materials. Cf. R. Filipović, ed., *Zagreb Conference on English Contrastive Projects*, Zagreb, Institute of Linguistics, 1971.

of contrastive analysis in teaching the target language then a representative corpus will offer much better and more versatile teaching material than the examples we use in a theoretical discussion to illustrate rules.

6.3. Although work on a contrastive project based on the T—G approach can begin without a corpus and be successful in contrasting equivalent rules in source and target languages, like PAKS, a corpus can be of great use in such projects in two directions: a) checking the functioning of the established rules, and b) furnishing examples by means of which new rules (that have not been established through intuition) can be formulated and investigated.

7.0. A well organized corpus represents the best linguistic text for the analysers, certainly better than some material gathered *ad hoc.*

It is impossible nowadays to make an analysis of some important sections of a language without exact date on distribution.

7.1. A corpus has advantages over informants[42] in giving information about distribution. An informant, for psychological reasons, gives us distribution for one person who is always under some pressure.

7.2. The distribution information which we get from grammars is not completely reliable either. A grammarian looks for examples to illustrate his theory. It is always dangerous for him to use only those examples he needs for his purposes in a certain stage of his analysis and to reject or neglect others. When using a corpus systematically this cannot happen.

7.3. A good corpus which is a large unit with organic continuity, and therefore a natural linguistic text which is also carefully structured at stylistic levels, can offer statistical reliability and representativeness.

8.0. From what I have said it is evident that adopting a corpus does not mean giving up theory. In the discussion at the Zagreb

[42] Several authors have expressed some doubts about the reliability of informants. Ilse Lehiste closes her article "Grammatical variability and the difference between native and non-native speakers" (in G. Nickel, ed., *Papers in Contrastive Linguistics,* Cambridge University Press 1971, pp. 69—74) with an interesting statement: "If there is so much variation among the native speakers and so much similarity between native and non-native speakers, the appeal to the native speaker's intuitive knowledge of grammaticality seems to lose much of its force". (p. 73).

Conference there was an interesting intervention. The speaker[43] emphasized that we had all agreed on the primacy of theory, and added in the form of a question that there are theories which are against adopting a corpus, and when a corpus is adopted theory changes but is not given up.

8.1. There is no contradiction between theory and corpus. Just the opposite! There is a strong interdependence between the two. What is different is the degree of their interdependence and the degree of applicability of the corpus in the contrastive analysis. In work with the structuralist approach and the translation method a corpus is more or less indispensable. If we use the generative approach a corpus is not needed in the initial stage. However, the further we go in our analysis the more useful a corpus can be.

9.0. There is no need to exclude native informants either. It is always useful and may even be necessary to check theoretical results on native speakers too. (They can be considered as a kind of a "living corpus".) This only means that we have double checking. In our work we do both. When we discussed some topics of our project and analysed them on our corpus we came to a point when we had to turn to our native informants before we could come to a final decision.

9.1. We are aware that even a very big corpus, like the Brown Corpus, can lack some items which we know by intuition ought to be discussed. Then only native informants can help. However, we never work only with informants. We bear in mind the fact that informants need not and cannot always be reliable; native speakers do not agree among themselves about what is grammatical.[44] Native informants are best used for additional checking after we have exhausted the help of the corpus.

10. Here are some points in conclusion: (1) A corpus cannot and should not replace theory, it should not come before theory nor instead of it; (2) No contrastive project can be regarded as complete before its results are verified and completed by means of a corpus; (3) Only a corpus can verify some doubtful cases of grammaticality; (4) We can assess the frequency and distribution of some forms only by means of a corpus; (5) Without a corpus we could not discuss the stylistic values, i. e. stylistic levels or registers, of some forms; (6) A corpus is indi-

[43] A. de Vincenz in R. Filipović (ed.), *Zagreb Conference on English Contrastive Projects, 7—9 December 1970. Papers and Discussion.* Zagreb, Institute of Linguistics, 1971, p. 151.

[44] Ilse Lehiste, *o. c.*, see note 42, p. 69.

spensable as one of the three components of the "contrastive mix" without which no contrastive analysis can be regarded as complete;[45] (7) Without a corpus it would be impossible to get a more or less exhaustive listing of all items that belong to a certain class, which is very important for contrastive analysis and its practical application.

[45] V. Ivir, "Generative and Taxonomic Procedures in Contrastive Analysis", in R. Filipović (ed.), *Zagreb Conference on English Contrastive Projects, 7—9 December 1970. Papers and Discussion*. Zagreb, Institute of Linguistics, 1971, p. 167.