Comparing cohesive devices: a corpus-based analysis of conjunctions in written and spoken learner discourse

Roumiana Blagoeva
Department of English and American Studies
Sofia University *St. Kliment Ohridski*

1.The aims of this paper are: to describe the process of developing two learner corpora of English at Sofia University; to mention some problems associated with the collection of the data as well as some of the applications of corpora of this type, and to present the findings from the research carried out so far.

2. The International Corpus of Learner English (ICLE) project, launched at the University of Louvain in Belgium to compile an international corpus of written learner language was joined by the Bulgarian team in 1996. To conform to the requirement for comparability all the teams participating in the project are collecting the same type of data, the only difference between the sub-corpora being the learners' mother tongue. At present the Bulgarian sub-corpus contains about 112 000 words, which is 55% of the total amount of data required, and work on it is still going on.

The data consist largely of argumentative essays written by Sofia University students at the beginning of their second academic year, so they can be described as adult advanced learners of English. Each of the 112 learners contributed about 1000 words on one or two of the essay topics suggested by the ICLE members.

In 1995, a complementary project was conceived in Louvain to compile a corpus of spoken learner language. The Louvain International Database of Spoken English Interlanguage (LINDSEI) project is the first of its kind and was soon joined by several other countries, including Bulgaria. In order to ensure comparability of the data each sub-corpus was to contain transcripts of 50 fifteen-minute interviews with non-native university students of English. The Bulgarian sub-corpus has already been compiled, and the transcription and keying-in of the data has been fully completed. The amount of words collected approaches 110 000 including the speech of the interviewers, who are native speakers of English.

3. A prime concern of this corpus creation activity is to collect data from a homogeneous population. Among the variables that need to be controlled are learning environment, age, mother tongue, stage of learning and nature of the task. The relevant biographical information about each contributing learner, such as years of English at school, stay in an English speaking country, knowledge of other foreign languages, use of reference tools, is encoded in a learner profile questionnaire which learners fill in(Granger, 1994: 26).

The group of learners who contributed to the Bulgarian sub-corpora is homogeneous in terms of all the variables listed above. The subjects are all second-year students of English and American Studies at Sofia University; aged 20 to 22; the reference tools used when writing the essays are monolingual dictionaries; their nationality, mother tongue and language spoken at home is Bulgarian with only 2 exceptions where one of the parents speaks Turkish at home; the medium of instruction at school and at university has been English and Bulgarian; and very few students (only 6) have spent from four months up to two years in an English speaking country.

Another essential consideration in developing the corpora is the representativity of the data to be collected. In compiling the two Bulgarian sub-corpora there was no selection of contributors on the basis of academic record or any personal preferences on the part of teachers and students. The essay writing as well as the interviews, their transcription and conversion into electronic form were incorporated in the students' written and oral assignments. In this way all Bulgarian students of English and American Studies at Sofia University could participate in the two projects and they did their best to fulfil the tasks. There are inevitable differences in the previous learning experience of the subjects, but I consider them an advantage rather than a hindrance because this fact practically excludes the influence of one and only one teaching strategy on the learners, thus, rendering the data representative of a much wider population than the one sampled. It should also be noted that since the ratio between male and female students at the Department is 1:10, the number of female participants

is much higher (90%) than the number of male ones. If the two sexes were to be represented equally the collection of data would have extended over a much longer period of time.

4. A learner corpus is very different from a native corpus because of the nature of the material collected. A native corpus contains data from a natural language and can be used on its own for the investigation of characteristic features of this language. A learner corpus presents evidence of an interlanguage; and an interlanguage, regardless of its stages of development, can only be a simulation of the natural language that is the target aimed at in the process of FLT. Therefore, any learner corpus would be of little value on its own, but it could be a useful tool for investigating a particular interlanguage when compared to a relevant native corpus. The choice of this native-speaker corpus is dependent on the aims of FLT. In preceding decades comparisons were largely carried out between learner language and the norm of the target language described in grammar books and dictionaries. Isolated examples of TL and IL were analysed with the purpose of finding erroneous structures and elements in the IL. So analyses tended to overlook the fact that learner language is not characterised only by misuse, but also by over/underuse of words, syntactic structures and discourse features. This is especially true for the highest levels of FLA, where errors are rare but we still feel that learners have not achieved near-native competence and learner production still differs from what native speakers would produce in similar situations. If the final goal of FLT/FLA is to achieve an ability to use the TL the way it is used by native speakers for the fulfillment of certain real-life tasks then a study of interlanguage will always need a corpus of authentic samples of the foreign language to compare with learner production.

5. For this reason two computerized native-speaker corpora of nearly the same amount of data were also developed at Sofia University. The written native language corpus is a collection of newspaper articles and essays on various topics; the spoken native language corpus contains transcripts of interviews, dialogues, announcements, extracts from radio programmes etc. All of the texts in these two corpora are used as teaching materials in- or outside class, and in test papers. The choice of text types and sources is justified by the fact that the most immediate contact Bulgarian students have with the English language is established through teaching and the media rather than personal contact with native speakers. Moreover, the students are trained to be specialists in English language and culture and on graduating should be able to communicate in English in a variety of situations as well-educated people. Therefore, a comparison of their production with the types of native corpora mentioned is relevant for the present study.

6. Having such electronic collections of natural textual data enables us to carry out research on a large scale and to explore characteristic features of interlanguages across a variety of backgrounds in a quick and reliable way. Instead of analysing isolated, invented sentences we now have the opportunity to explore language in use with its purposes and functions in mind.

> "… for us the actual text, not the invented sentence, must be the essential linguistic unit … In the coming millennium, this prospect can now finally be documented and clarified by working with very large corpora of authentic texts, whereby we can hope to uncover some of the vital and delicate missing links between 'language' and 'text'." (Beaugrande, 2000)

As noted by Halliday (Halliday, 1976:2), "a text does not consist of sentences; it is realized by, or encoded in sentences". Hence, only the study of longer stretches of discourse can give an insight into the resources that English has for creating texture.

7. Some of the first searches applied to the Bulgarian sub-corpora are connected to the use of conjunctive elements by the learners. Conjunctions seem to be the most explicit way of establishing relations in a text since they indicate how what is to follow is systematically connected to what has gone before in the text (Halliday, 1976: 227). They are a means that text producers use "to exert control over how relations are recovered and set up by the receivers" (Beaugrande, 1983: 74). Conjunctive elements with their intrinsic function to signal relations in a text are rather different from the rest of the lexicon in a language: they are relatively independent of context in the sense that we can expect them to be present in a text no matter what the particular topic of this text is. The study of their use by foreign learners of English can provide the first step to the understanding of the ideas that learners have about text structure and to the investigation of their ability to construct a text.

8. The present analysis examines fifty words and phrases that, according to Halliday's (Halliday: 242-243) classification of conjunctive relations express additive, adversative, causal and temporal relations, in a text (*and*, being one of the most frequently used additive conjunctions, is the subject of

a separate study and is excluded from this analysis). Using WordSmith Tools (Scott, 1997), concordances were produced for the fifty conjunctions in each of the four corpora. In Corpus 3, containing the learner spoken data, the interviewers' words were manually excluded and only the interviewees' speech was taken into account. This slightly reduced the size of this corpus so the frequency of occurrence of each item was calculated as a percentage of the total number of tokens in the corpora. The results of this first search are summarised in Table 1.

Number of conjunctions studied	Percentage of occurrences in each corpus			
	Corpus 1 Learner written 100 000 words	Corpus 2 Native written 100 000 words	Corpus 3 Learner spoken 70 000 words	Corpus 4 Native spoken 100 000 words
50	3,627	1,894	4,328	3,364

Table 1. Frequency of occurrence of 50 cohesive devices in four corpora

8.1. There are two striking facts that these figures reveal: first, the greater use of connectors by both native speakers and learners in speaking than in writing; and, second, a clear overuse of conjunctions by learners in written and spoken production respectively in comparison with native speakers.

8.1.1. A number of scholars studying spoken language (e.g. Labov, 1972; Chafe, 1979; Cicourel, 1981; Goffman, 1981; Biber, 1995) have reached the conclusion that conjuncts are more formal and therefore more typical of written rather than spoken language because the speaker is usually less explicit than the writer. It is true that the speaker can resort to a number of means of expression such as gestures, posture, tone, etc., but it is equally true that in most situations the speaker is under the pressure of time-limitations as he observes his interlocutor and has to respond quickly to his reactions, sometimes modifying what he is saying and making it clearer and more concise. The writer, on the contrary, can construct a text at leisure, can spend more time on the choice of particular words and syntactic structures, or can go over the whole text and edit it in many different ways. This allows the writer to avoid certain unnecessary repetitions (including the repetitions of formal connectors) and to find other means of structuring his text in a logical way.

Whatever the reason, I am fully aware of the fact that working with corpora of 100 000 words each cannot give a detailed picture of the existing state of written and spoken language. What this study presents is based only on the observations of the raw data from the discourse stretches in the particular corpora under investigation. I believe that the extension of the data and further analyses and collaboration with the other teams may explain this phenomenon.

8.1.2. In trying to provide reasons for each deviant use of the target language by learners we most often turn to the factors influencing FLA and the psycholinguistic processes central to second language learning as determined by Selinker (1972: 37).

The overuse of conjunctions by learners of English could be due to some teaching/learning strategies. In most classrooms, and more specifically in Bulgaria, speaking and writing are distinguished as different skills and are trained separately and very often independently. Teaching materials used in English classes and in the special writing courses place too great an emphasis on text structure and the importance of formal connectors in general while at the same time leave very little space for other ways of achieving coherence. This may lead learners to overgeneralization of some TL rules and may bring them to the conclusion that a well-written and logically structured text can be produced only if such language items are abundantly present in it. Such an idea results in a tendency on the part of learners to "paste" connectors mechanically to the texts they produce in an attempt to give them some "shape" (Blagoeva, 2000: in press).

By contrast, the development of speaking skills usually goes along the lines of free discussions on suggested topics. Little instruction is given on how to speak coherently and to participate effectively in the act of communication. Greater attention is paid to grammatical accuracy and pronunciation as well as to knowledge connected with the topics discussed. It should be noted, however, that the contributors to the two learner corpora are highly influenced by long and constant exposure to written forms of language. Since our Department offers simultaneous courses in Linguistics and English and

American Literature the students' work is mostly directed towards reading and writing tasks. This probably accounts for the overuse of conjunctive elements and the more formal register they use in their speech.

9. The overall overuse of conjunctions by the learners could be misleading if concrete examples were not examined in greater detail. Table 2 presents the occurrences of the top 15 connectors in descending order of frequency.

Top 15 Conjunctions	Percentage of occurrences in each corpus			
	Corpus 1 Learner written 100 000 words	Corpus 2 Native written 100 000 words	Corpus 3 Learner spoken 70 000words	Corpus 4 Native spoken 100 000 words
But	0,632	0,426	0,988	0,632
Because	0,203	0,062	0,534	0,238
So	0,132	0,033	0,432	0,429
However	0,108	0,032	0	0,012
Then	0,107	0,090	0,234	0,299
For instance \ For example	0,098	0,017	0,062	0,058
Though \ Although	0,091	0,077	0,058	0,072
Of course	0,066	0,020	0,074	0,107
Therefore	0,056	0	0	0,020
Thus	0,050	0,002	0,004	0
Rather	0,040	0,016	0,020	0,025
Actually	0,035	0,013	0,235	0,135
In fact	0,033	0,014	0,122	0,059
Yet	0,022	0,015	0,003	0,010
I mean	0,006	0	0,245	0,144

Table 2. Frequency of occurrence of the top 15 connectors

9.1. Even a cursory glance at the first two columns shows that the greater number of instances of conjunctions in the written learner production are almost evenly distributed among all the 15 most frequently used connectors. In a recent paper Blagoeva (2000: in press) also reports that the results obtained from the Bulgarian written learner data confirm the findings made by Granger (1994: 27-28) about the overuse of the same connectors in the writing of French learners. (For further details see Blagoeva, 2000: in press)

9.2 The data extracted from the spoken corpus, however, demonstrates distribution of connectors in the speech of the learners different from that of the native speakers. Two of the connectors, *so* and *then*, have nearly the same ratio in Corpus 3 and Corpus 4 probably because they are also perceived by the learners as pause fillers. *But, because, in fact, I mean, actually* seem to be favourites of Bulgarian speakers of English and it is quite understandable if NL interference as a factor influencing learners' choices of words is considered. The English *in fact* can introduce "a contradiction or an opinion which is different from something that has just been said" (COBUILD,1994) and is therefore classified as adversative, contrastive conjunction (Halliday, 1976: 242-243). The Bulgarian translation equivalent of *in fact*, *vsăštnost*, means only *in reality* and is used to introduce some clarification or to add details to a previously mentioned statement (Bulgarian Language Dictionary, 1994). In this way it functions rather as an additive, and since this is the more frequent relation in a text it is quite natural to find more instances of this connector in the Bulgarian learner speech.

The other conjunctions that show greater frequency in the Bulgarian-English interlanguage have functions similar to those of their translation equivalents in the NL of the learners. *But* in English and *no* in Bulgarian, for example, have the same contrastive adversative function; *because* and *zaštoto* in English and Bulgarian respectively "introduce the reason for a statement or the answer to a 'why' question" and are both causal conjunctions. There seems to be no place for confusion here and it is true that no instances of misuse were encountered. Still NL interference could work not only because of formal differences between languages but also through the transfer of different speaking or writing habits in the mother tongue due to some cultural differences. One hypothesis at this point is that Bulgarians tend to be more explicit when stating reasons or objections. Yet such a conclusion could be

confirmed only after comparisons with relevant Bulgarian native corpora, which are still being compiled at Sofia University.

At the same time several connectors on the list point to a tendency for underuse in learner speech. Obviously, some language items are used at the expense of others whenever students do not feel confident enough in their knowledge of the foreign language.

10. One could argue that the higher or lower frequency of formal connectors in learner language, as long as they are used correctly, may not lead to serious communication breakdowns. Still, it is my view that it could interfere with a receiver's comprehension of a text and could contribute to the artificiality of learner English. At an advanced stage of FLA students should be made aware that they tend to stick to some language structures and should be encouraged to turn to other means of achieving cohesion.

Another salient point is connected with the differences between spoken and written language and the choice of teaching materials for the development of speaking and writing skills. By revealing characteristic features of learner language, corpus analysis studies offer ways of diagnosing the true learners' needs for the different purposes of communication. The results of such studies can turn the attention of teachers to the fact that *speaking* is not equivalent to *mere talking* but is a special skill that can be trained in a systematic way. Naturally, further corpus-based investigation of other discourse features is likely to be the way forward to developing learner resources that will bring interlanguages closer to the kind of language used by native speakers of English.

References:

Andreichin L, et al 1994 *Bălgarski Tălkoven Rečnik [Bulgarian Language Dictionary].* Sofia: Nauka i izkustvo,.

Beaugrande R 2000 Text linguistics at the millennium: Corpus data and missing links.*Text*, 20, 2000.

Beaugrande R, Dressler W 1983 *Introduction to text linguistics.* London and New York: Longman.

Biber D 1995 *Variarion across speech and writing.* Cambridge; Cambridge University Press.

Blagoeva R 2000 Comparing cohesive devices: conjunctions and other cohesive relations and their place in the Bulgarian-English interlanguage. *Paper presented at Third international conference for research in European studies*, Veliko Turnovo, Bulgaria.

Chafe W L 1979 The flow of thought and the flow of language in (ed) T. Givon.

Cicourel A 1981 Language and the structure of belief in medical communication in (eds) B. Sigurd and J. Startvik, *Proceedings of AILA 81 Studia Linguistica* 5: 71-85.

Goffman E 1981 *Forms of talk.* Oxford: Basil Blackwell.

Granger S 1994 The learner corpus: a revolution in applied linguistics. *English Today* 39: 25-29.

Halliday M A K, Hasan R 1976 *Cohesion in English*. London and New York, Longman.

Labov W 1972 *Sociolinguistic Patterns Philadelphia*: University of Pensilvania Press.

Scott, M. 1997. Wordsmith Tools. Version 2.Oxford: Oxford University.

Selinker L 1972. Interlanguage. *International Review of Applied Linguistics* 10: 209-31.

Sinclair J, ed. in chief. 1994. *Collins Cobuild English Language Dictionary*. London: HarperCollins Publishers.