BILINGUAL CORPORA AS A PLATFORM FOR CROSS-LINGUISTIC TREEBANK DEVELOPMENT[1]

Tzvetomira Venkova

Sofia University
venkova@ttm.bg

0. Introduction

The paper aims at exploring the opportunity of extending an existing treebank of one language to covering other languages by way of analysis, based on two specific principles: corpus orientation and partiality. The suggested approach comes as an attempt for a possible alternative to the purely theory-based approaches, where grammar theories for one language are transferred into the other language and subsequently examples are sought to support them. I do not mean, however, to discard the need of theory base but rather to try to motivate the integration of theory and empirical evidence from both languages at an early stage of treebank development. When the process of seeking theory comparability of languages is being carried out in close connection with corpus reference, the resultant treebank is supposed to be more sensitive to the specifics of the second language and allows flexibility in theory moderation at the very beginning of treebank construction. Moreover, if the treebank is not considered as an end in itself but is conceived in regard to potential applications to real dialogue systems and communicative language teaching, where authentic speech is involved, the empirical evidence of the corpus and respectively of the retrieval collections based on it, is essential.

1. Bilingual comparable retrieval collection (BCRC)

The paper introduces the notion of *bilingual comparable retrieval collection*[2] (BCRC) to term the intermediate step between an unannotated comparable corpus and a fully finished, consistent bilingual treebank. BCRC originates from the corpora of the two languages but is not a corpus proper since it is a selection, based on certain criteria, and as such, it violates *naturalness* (Sinclair

[1]Research on this paper was supported by a DAAD grant at the Seminar fuer Sprachwissenschaft, Tuebingen University. I am very grateful to Erhard Hinrichs for his professional help and advice and also to Sandra Kuebler for her help with the *The Tuebingen Verbmobil Treebank of English*. Thanks also to Petyo Byankov and Frank Richter for software assistance and comments.

[2]The term *retrieval collection* was suggested to me by John Sinclair in personal communication.

1984) as a basic feature of authentic corpora. BCRC is related to the notion of treebank as the selection of its content is based on lexical, morphological and syntactic criteria but as a general design it has a looser structure, various types of information and is subject to change. In a way, it can be regarded as a test-bed for a treebank. BCRC is also related to the notion of *test suite* (Lehman et al. 1996) but it differs in its procedural aspect and deliberate closeness to corpus.

Compared to a concordance, BCRC is more complicated as the search criteria for its compilation are not necessarily strings from the text but can be semantic, syntactic, etc. Its construction involves human labour, as it can not be compiled in a fully automatic way, but the time invested in it is further compensated, as once created BCRC can stay as a supplement to the corpus and can be used for various purposes. What makes it different from individual random search results of a linguist, is that it is supposed to have clear structure and function, so that other people can use it.

The construction of BCRCs as separate entities might raise an objection that the number of BCRCs created for one corpus can be unlimited and in a way uncontrollable. They, however, are procedurally defined as steps to a treebank, and hence the usefulness of each BCRC depends on the needs and aims of the particular treebank. And as they would be more or less grouped around the basic grammar and lexicon topics, they can be used also for purposes beyond the treebank, such as banks for language teaching, for example.

Apart from corpus orientation, another feature of BCRC is its partiality. Naturally, the ideal case would be to build an entire treebank at one step but it is rather difficult, especially when less widespread languages such as Bulgarian are concerned. Thus BCRC is an attempt to make at least a partial step to build a treebank. On the other hand, it is clear that partial approaches can lead to inconsistency in a larger structure. In an attempt to avoid this problem and make BCRC as comparable as possible, the suggested analysis of the BCRC of time clauses has been kept at the lexical-syntactic level of the tree structure and it follows basic syntactic principles of more or less general acceptance.

Since the use of the notion *comparable* in characterising a bilingual corpus might rise some terminological misunderstanding, I am briefly discussing it here. The literature survey shows that at least three different term pairs define one and the same typological distinction of bilingual corpora: *translation* - *parallel* (Ajmer & Altenberg 1996:13, Granger 1996:38), *parallel* - *comparable* (Baker 1993:248, Erjavec et al 1996) and *translation* - *comparable* (Granger 2001). The observations show that the distinction *parallel* (for a corpus consisting of original texts in one language and their translations into the other language) and *comparable* (for a bilingual corpus of original texts with comparable genre or text type in the two languages) seems to be most widespread in the research literature and is in common use in discussions concerning bilingual corpora. For this reason, it has also been adopted in this paper - the bilingual corpus of spoken English and Bulgarian is termed a *comparable bilingual corpus* and the retrieval collection, based on it, is called respectively a *comparable retrieval collection.*

The particular BCRC, discussed in the paper, contains temporal relations, namely time clauses in the corpora of English and Bulgarian. Temporal relations have been chosen not only because they are connected with the general orientation of The Tuebingen Verbmobil Treebank of English but also because their expression in colloquial speech has its specific character which is a challenge for

more detailed research from theoretical point of view. Due to its content, this BCRC is called *'Time clauses'*

The starting point of *BCRC 'Time clauses'* was *The Tuebingen Verbmobil Treebank of English* (cf. Hinrichs et al. 2000 a, b, Kordoni 2000) and in the course of the research has proved to be a reliable base for further research and investigation.

The BCRC TIME CLAUSES carries on the tradition of the research in the field of bilingual corpora concerning Bulgarian, such as *Linguist Workbench* (cf. Stambolieva 1996) and *Multext - East* (cf. Erjavec et al 1996) and in its turn, it pays more attention to the characteristics of colloquial speech and namely to its contrastive temporal characteristics in Bulgarian and English.

2. Sources of the comparable retrieval collection

Two corpora - of English and Bulgarian respectively - have been used as sources of *BCRC 'TIME CLAUSES''*. Each of them contains approximately 45 000 word-forms.

The corpus of English colloquial speech is based on collection CD8 of *The Tuebingen Verbmobil Treebank of English.* This bank is one of the products of the international project Verbmobil, developed with regard to the machine translation (cf. Wahlster 2001, Hinrichs et al. 2000 a, b). *The Tuebingen Verbmobil Treebank of English* has been used in the development of the English-Bulgarian comparable corpus in various aspects - as transcripts of English dialogues, as tree-structures, corresponding to each sentence of the transcripts, as a search and statistics tool and as a methodological source.

The corpus of Bulgarian colloquial speech contains dialogues transcripts. It is part of a larger corpus, collected by Tz. Nikolova on paper as a source for her Frequency dictionary of Bulgarian colloquial speech, cf. Nikolova 1987. With the kind permission of Tz. Nikolova, I have typed in and designed as a corpus a considerable part of it -50 000 word-forms, cf. Venkova 1996:264. This corpus is publicly available in Internet at *http://www.hf.vio.no/easteur-orient/bulg/mat/index.html/Nikolova* thanks to the initiative of Kjetil Ra Hauge from the University of Oslo.

The two corpora, presented above, have been sources of the respective monolingual retrieval collections of time clauses, united further into a bilingual retrieval collection.

Although, in regard to thematic scope, the English corpus is more restricted, since the conversations there concern mainly appointment negotiations and scheduling appointments, it does not significantly affect the corpus analysis. In view of the advantages of bilingual corpora, presented in Lauridsen 1996:63 the source corpora have the necessary potential for contrastive studies.

3. Stages of compiling the *BCRC 'TIME CLAUSES''*

The compilation of *BCRC 'TIME CLAUSES'* involves several stages, which have been developed by computer-aided procedures.

3.1. Preliminary research

As a preliminary research, the structural-semantic aspects of time clauses in English and Bulgarian have been investigated and an itinerary of the basic operational criteria, concerning the structural varieties, semantic guidelines and borderline cases have been compiled. The main theoretical sources have been included in the paper bibliography.

3.2. Concordance of time conjunctions

The fastest way of selecting time clauses is to perform automatic search of time conjunctions, connecting them. This has been done by giving each conjunction as a key word and producing a concordance of all conjunction occurrences within the respective complex sentences, regarded as minimal context. Since time conjunctions are a very limited number and moreover they are non-inflected part of speech, the searching procedure has been performed technically very easily by an automatic concordancer. The following conjunctions have been employed as key words: (Bulg.) *dokato, dokle, dokogato, dorde, dordeto, kato, koga, kogato, kogato i da, otkak, otkakto, otkoga, otkogato, predi da, sled kato, shtom, shtom kato*; (Engl.) *after, as, as early/ long/ often/ soon as, at the very moment, before, every time (that), during the period (that), no sooner than, now that, once, since, so long as, throughout (that), till, until, when, whenever, while*.

Time conjunctions in the English corpus have been extracted by the search function built in the program package of *The Tuebingen Verbmobil Treebank of English*. The program identifies all occurrences of a conjunction and presents them together with the corresponding complex sentence. After that, each of the occurrences, together with its sentence context, has been saved as a separate file. The concordance of Bulgarian time conjunctions has been produced by a macros.

Actually, the context of the sentence in of *The Tuebingen Verbmobil Treebank of English* comes as a parsed sentence with a tree structure. The tag set, however, does not identify time clauses and time conjunctions. Therefore the *BCRC 'TIME CLAUSES'* analysis can be regarded as a step towards possible further specification of the tag set and also as a test-bed for the existing more general tag set.

The English part of the concordance contains 484 occurrences of the above listed conjunctions, and the Bulgarian part, respectively - 449 occurrences.

3.3. Semantic analysis of time conjunctions concordance

Automatic concordancing, however, was only the first stage of time clause extraction, since it was not precise enough to select a group of sentences only according to the type of the conjunction, connecting their clauses.

The first group of problems comes by the fact that many non-inflected words are polyfunctional, thus, some of them, apart from being conjunctions, have a number of other functions. For example, *before* (Engl.) and its equivalent *predi* (Bulg.) can function as a subordinate conjunction (cf. 1 a, b below*)*, a preposition (cf. 2 a, b below), and an adverb (cf. 3 a, b below) and, in addition, *before* can also be an adjective in English (cf. 4 below). This polyfunctionality is illustrated by sentences 1 – 4 below, all of them taken from the source corpora:

(1a) *That way, we have four hours before I have anything to do.* (Engl.) *{TVTE, CD8, sentno 1370}.*

(1b) *Az, chistihme, predi da dojde svekyrva mi, nali, osnovno chistihme bibliotekata, vsichko.* (Bulg.)*{CBCS, R16, 018}.*
'Me, [we] cleaned before my mother-in-law came, well, cleaned the library thoroughly, everything'.

(2a) *That sounds good, are you available just before noon?* (Engl.) *{TVTE, CD8, sentno 408}.*

(2b) *Sinyt zamina predi mene na uchilishte.* (Bulg.) *{CBCS, R05, 081}*
'My son left for school before me'

(3a) *You ever been there before?* (Engl.) *{TVTE, CD8, sentno 2551}*

(3b) *I predi stavashe kym shest, shest i polovina, kukurjak.* (Bulg.) *{CBCS,R06, 171}*
'And before [he] woke up about six, six thirty, early riser'

(4) *I will be out of town the weekend before.* (Engl.) *{TVTE, CD8, sentno 1547}*

Another example of such polyfunctionality is English *when,* which can function as a subordinating conjunction (1a above) or as an interrogative adverb (cf. 5 below):

(5) *When will you be back?* (Engl.) *{TVTE, CD8, sentno 123}*

In standard Bulgarian these two functions are expressed by two different words: *koga* and *kogato*. In spoken Bulgarian, namely in West Bulgarian speech, however, *koga* is often used in both functions (cf. 6 and 7 below):

(6) *I ne moga, koga ima vjatyr, az ne moga da takova, da spja.* (Bulg.) *{CBCS,R06, 143}*
'I can't, when there is wind, I can't sleep'.

(7) *Koga pristigna avtobusyt ot Govedarci?* (Bulg.) *{CBCS,R11, 034}*
'When did the bus from Govedarci arrive?

Since such cases are numerous in the Bulgarian corpus, they have been considered in the disambiguation process.

The second type of problems concern cases in which the function of the keyword is that of conjunction but this conjunction introduces different types of subordinate clauses, for example *when* in English and *koga 'when'* in Bulgarian can introduce time clauses (cf. 8 a, b below), object clauses (cf. 9 a, b below) and relative clauses (cf. 10 a, b below), which causes a different type of ambiguity, cf *when* in 8-10.

(8a) *If you want to meet when I am just coming back from vacation, we could meet on the seventeenth, I suppose.* (Engl.) *{TVTE, CD8, sentno 1386}*

(8b) *Pishe bezuprechno na mashina, kogato podnasja kafe, vsichki sa v zahlas.* (Bulg.) *{CBCS, R03, 143}*
'[She] types faultlessly, when [she] serves coffee, everyone is dazed.'

(9a) *I do not know when you will be available.* (Engl.) *{TVTE, CD8, sentno 2228}*

(9b) *I ela utre da takova, da ti kazhem koga shte zapochnat zanjatijata.* (Bulg.) *{CBCS, R04, 110}*
'Come tomorrow to, well, to tell you when the classes begin.'

(10a) *Ahm, you just tell me a time when you are free, and I should be able to take care of that.* (Engl.)*{TVTE, CD8, sentno 1351}*

(10b) *I vyv denja, kogato az se namiram tuka, za da si vzema chasa, toj mi vdiga skandal.* (Bulg.) *{CBCS, R21, 030}*
'And in the day when I am here to take my lesson, he kicks up a row.'

Because of the above shown conjunction polyfunctionality and ambiguity, it was necessary to perform semantic-grammatical analysis to all automatically selected sentences as a next step towards the precise extraction of time clauses.

The parallel between the conjunction word entries in the two languages and also the reference to some theoretic works and dictionaries, such as *Quirk et al 1972, Oxford Advanced Learner's Dictionary 1989, Collins Cobuild Dictionary 1987, GSBKE 1989, Penchev 1993, Maldjieva 1995, Tisheva 2000, Ra Hauge 1999, Bylgarski tylkoven rechnik 1995* have been very useful in the analysing process.

As a result of this syntactic-semantic analysis, it has become possible to select only those complex sentences in which a particular conjunction introduces a time clause. After their selection, the sentences have been marked with the appropriate tag, thus making way for all basic sorts of automatic processing such as recognition, extraction and sorting.

Some specific difficulties to the analysis have been caused by the colloquial nature of the corpus. There are many sentences that can not be licensed by structures extracted from fiction corpora. Such typical speech phenomena as unfinished sentences, repetition, ellipsis, strategy changes, etc. described by many authors, cf. for example Angelova 1994, Zemskaja 1987, are very common in the comparable speech corpus.

Another challenging issue for the analysis were some differences in the grammar traditions. Similar language phenomena in the two languages are interpreted by different terminology and in different aspects in Bulgarian and English linguistic literature, which motivated discussion and careful comparison in the course of the analysis. On the other hand, such a contrastive perspective is very useful for the distinction of some specific features of the investigated structures in each language.

3.4. Structuring and integration of the monolingual retrieval collections into *BCRC 'TIME CLAUSES'*.

After the semantic analysis and the final selection of time clauses, they have been further processed in two steps. As a first step, they have been grouped and structured as two monolingual collections for each language and, after that, they have been integrated into a bilingual comparable retrieval collection.

The structure of each monolingual retrieval collection has been developed in regard to two principal considerations: firstly, providing maximal facilitation for automatic search and further integration in NLP tools; secondly, focusing on clear grammar and lexical patterns which need to be singled out and which should be easy to trace for research and teaching purposes. Each sentence has been given an index, indicating its exact position in the source corpus. For the English sentences the index refers to the running number of the sentence as numbered in CD8 *of The Tuebingen VERBMOBIL Treebank of English* and for the Bulgarian corpus respectively - the sample and the line number in which the time conjunction occurs, according to the corpus numbering described in Venkova 1996. The sentences can be retrieved according to particular

criteria and they can easily be re-grouped according to other criteria. Cross-linguistic references are also possible due to the indexing.

Next, the monolingual collections have been integrated into the *BCRC 'TIME CLAUSES'* and it is further structured according to the functional equivalents of the subordinating conjunctions. Each functional group can be further subdivided according to semantic, structural or lexical criteria.

The ***BCRC 'TIME CLAUSES'*** contents have also been statistically processed and the collected data allows retrieval of frequency profiles that can possibly be used in programs for statistical natural language processing.

3.5. Supplementary banks

The sentences included in the automatic concordance but rejected in the semantic analysis, have also been sorted as accompanying corpus collection since they can also be very useful in any kind of research or teaching activities, concerning the distinction of time clauses.

4. *BCRC 'TIME CLAUSES'* as a source for automatic retrieval of linguistic information

4.1. Automatic search according to a particular conjunction

The *BCRC 'TIME CLAUSES'* can give an opportunity for searching and extracting time clauses, if the subordinating conjunction is given as a keyword. This is a fast procedure, which produces as a result a concordance with the corresponding frequency information. As this procedure is based on the results of the semantic analysis described above, it will include only those conjunction occurrences that have a temporal function.

4.2. Automatic search of colloquial and regional variants

Bulgarian part of the retrieval collection has been designed so that some regional and colloquial variants of time clauses can be displayed. Such is the use of *koga*, in stead of *kogato*, in the dialogues recorded in Sofia (see 6 above) and *kogat'*, in stead of *kogato*, in the Central Balkan speech, e.g.:

(11) *Kogat' si doide basta ti, ste mu kazha.* (Bulg.)
'*When your father comes home, I will tell him*'.

This provides a base for analysis regarding a broader treatment of colloquial speech patterns. It is also possible that these variants can be extracted in connection to each other or to the standard structures.

4.3. Extraction of functional equivalents

BCRC 'TIME CLAUSES' gives an opportunity for concurrent extraction of functional equivalents of time clauses in the two languages.

5. Conclusions

The suggested procedure for compiling a *BCRC 'TIME CLAUSES'* could be used as a prototype for compiling comparable retrieval collections with wider scope and can be further developed into a syntactic treebank in which the investigated syntactic phenomena can be treated in a more comprehensive way.

Corpus sources:

TVTE: *The Tuebingen Verbmobil Treebank of English*, CD8

CBCS: *The Corpus of Bulgarian Colloquial Speech: Nikolova*

Bibliography:

Aijmer, K. &Altenberg, B. (1996) Introduction. In: *Languages in Contrast*, Lundt Studies in English 88, S. Baeckman & Svartvik, J. (eds.), Lund University Press, pp. 10-16.

Angelova, I. (1994) *Syntacsis na bylgarskata razgovorna rech [v sypostavkas ruski, cheshki i polski ezik]*, Universitetsko izdatelstvo"Sv. Kliment Ohridski, Sofia.

Baker, M. (1993) Corpus Linguistics and Translation Studies: Implications and Applications. In: *Text and Technology. In Honour of John Sinclair*, ed. by M. Baker, G. Francis & E. Tognini-Bonelli, pp. 233-250.

***Bylgarski tylkoven rechnik* (1995)** L. Andrejchin i dr., Nauka i izkustvo, Sofia.

***Collins Cobuild English Language Dictionary* (1987)** John Sinclair (ed.), Collins, London and Glasgow.

Erjavec, T. et al (1996) Erjavec, T., Idle, N. Petkevic, V and Veronis, J. Multext-East: Multilingual Text Tools and Corpora for Central and East European Languages. In: *Proceedings of the First European TELRI Seminar: Language Resources for Language Technology*, pp. 87-98.

Granger, S. (1996) From CA to CIA and Back: An Integrated Approach to Computerized Bilingual and Learner Corpora. In: *Languages in Contrast*, Lundt Studies in English 88, S. Baeckman & Svartvik, J. (eds.), Lund University Press, pp. 37-51.

Granger, S. (2001) Corpora in Contrastive Linguistics, Translation Studies and Cross-linguistic NLP Applications, Paper presented at 6th TELRI Seminar, Bansko, Bulgaria, 9-11 November 2001.

GSBKE (1983) *Gramatika na syvremennija bylgarski knizhoven ezik, t. I - III*, Izdatelstvo na BAN, Sofia.

Hinrichs et al (2000a) Hinrichs, E., Kuebler**,** S., Kordoni, V., and Mueller, Fr., Robust Chunk Parsing for Spontaneous Speech. In: *Verbmobil: Foundations of Speech to-Speech Translation"*, German Research Centre for Artificial Intelligence, DFKI, Saarbruecken, Germany 2000, pp 163-183.

Hinrichs et al (2000b) Hinrichs, E., Bartels, J., Kawata, Y., Kordoni., V, Telljohann, H., The Tuebingen Treebanks for Spoken German, English and Japanese, In: *Verbmobil: Foundations of Speech-to-Speech Translation"*, German Research Centre for Artificial Intelligence, DFKI, Saarbruecken, Germany 2000, pp. 552-557.

Lauridsen, K. (1996) Text Corpora and Contrastive Linguistics: Which Type of Corpus For Which Type of Analysis? In: *Languages in Contrast*, Lund Studies in English 88, S. Baeckman & Svartvik, J. (eds.), Lund University Press, pp. 63-73.

Lehman et al (1966) TSNLP-test Suites for Natural Language Processing. In: *Proceedings of COLING 96*, Copenhagen 1996.

Maldjieva, V. (1995) *Non-inflected Parts of Speech in the Slavonic languages. Syntactic characteristics.* Wydawnictwo Energia, Warszawa.

Nikolova, Tz. (1987) *Chestoten rechnik na bylgatskata razgovorna rech.* Nauka i izkustvo, Sofia.

Oxford Advanced Learner's Dictionary of Current English **(1987)** A.S. Hornby, Oxford University Press.

Penchev, I. (1993). *Bylgarski sintaksis. Upravlenie i svyrzvane,* Plovdivsko universitetsko izdatelstvo, Plovdiv.

Quirk, R. et al (1972), Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. *A Grammar of Contemporary English*. Longman.

Ra Hauge, K. (1999) A Short Grammar of Contemporary Bulgarian, Slavica Publishers, Bloomington, Indiana.

Sinclair, J.M. (1984) 'Naturalness' in Language', In: Corpus linguistics, Aarts, J. and Meijs, W. (eds.), Amsterdam: Rodopi.

Stambolieva, M. (1996) A Linguist's Workbench. In: *Papers from The First Conference on Formal Approaches to South Slavic Languages (Plovdiv October 1995), University of Trondheim, Working Papers in Linguistics 28,* pp. 293-301.

Tisheva, J. (2000) *Modeli za interpretacija na slozhnoto izrechenie v bylgarskija ezik,* IK "SEMA RSH", Sofia.

Venkova, Tz. (1996) Kompjutyren konkordans na dumata *da* v razgovornata rech. V: *Problemi na sociolingvistikata 5,* M. Videnov, A. Angelov, Kr. Aleksova, P. Sotirov (systaviteli), Mezhdunarodno sociolingvistichesko druzhestvo, Sofia, pp. 263-266.

Wahlster, W. (2000)(ed.) *Verbmobil: Foundations of Speech-to-Speech Translation*. German Research Centre for Artificial Intelligence, (DFKI), Saarbruecken, Germany.

Zemskaja, E. (1987) *Rysskaja razgovornaja rech: lingvisticheskii analiz i problemyi obuchenija,* Moskva.