Wiener Slawistischer Almanach 52 (2003) 7-17

Angel G. Angelov

LEXICAL COHESION IN BULGARIAN LEGAL DISCOURSE

1. Method and Tasks

This paper describes the goals and the algorithm of a study entitled "Lexical Cohesion in Bulgarian Legal Discourse" and provides an overview of its second part, "Lexical Collocations in Bulgarian Legal Discourse." The study in question follows certain recent trends in the field of Natural Language Processing (NLP) and Computational Lexicography. The problems of computational lexicography, in turn, are immediately related to the analysis of corpora, which are linear structures and where the identification of the universal features, on the one hand, and the systemic modeling, on the other hand, are the result of a "pleasant" processing of empirical data.

1.1. As we know, classical *text linguistics* in the school of Halliday and Hasan has evolved into *corpus linguistics*, which has provided a fresh perspective on the at least one-hundred-year-old linguistic problem formulated by F. de Saussure: the question of the syntagmatic and paradigmatic levels of language. Computational analysis provides an easy and effective way of testing paradigmatic structures in the syntagm and vice versa, i.e. it provides an opportunity for extracting paradigmatic relationships from the linear character of speech.

Contemporary linguistics distinguishes between *semantic relations* and *formal relations*. *Semantic relations* are also discussed by text linguists, followers of Halliday and Hasan (1976), as well as de Beaugrande and Dressler (1981), van Dijk (1981), etc., but they are also discussed by Lyons (1977), Cruse (1986), Wierzbicka (1999), Miller and Fellbaum (1992).[1] The question and task of the study under consideration is to link the two types of semantic relations – horizontal and vertical. The analysis of *lexical collocations* appears to be very important for these semantic relations (Firth 1957, Halliday 1966, Halliday and Hasan 1976), although later both Halliday and Hasan abandoned this term (Hasan 1984, Halliday and Hasan 1986). Nevertheless, this approach – to

[1] G. Miller is the leader of the project on the so-called Online Lexical Database for English (WordNet) at Princeton University.

observe the combinations of the words and to study their variations on a formal as well as on a mental level – has been developed far in the works of John Sinclair (1991).[2]

1.2. The study in question is based on the computational analysis of a corpus of 500,000 words, drawn entirely from the genre of Bulgarian Statutory Texts. The corpus includes the Constitution of the Republic of Bulgaria, the Code of Civil Procedure, the Code of Criminal Procedure, the Family Code, the Sea Areas, Inland Waterways and Ports Act, the Road Traffic Act, etc.: a total of 49 current Bulgarian laws which would cover 800 pages if printed out. In terms of content, the study thus belongs to the field of *forensic linguistics*, which is devoted to legal discourse analysis. Forensic linguistics is also a new and borderline discipline of linguistics.[3]

Forensic linguistics (cf. Knifka 1990) is mainly designed to contribute to the objective character of legal and court proceedings – an extremely important and humane task. The study presented in this paper, I must admit, tends to "exploit" statutory texts, like any coherent texts, for the purpose of testing and applying statistical and computational procedures. However, as we shall see from the analysis, there are no texts devoid of meaning – Shcherba's *glokaya kuzdra* ... is applicable to syntactic but not to pragmatic relations, i.e. it does not apply to text structures. Semantic relations at both the syntagmatic and paradigmatic levels embed the text within reality, whereas reality, as Hasan shows, determines intextual variability. The same conclusions are also drawn by Paul Ricoeur (1986) in his study of the relations between text and reality, as well as by Halliday and Hasan in their 1986 book defining the term *context configuration*.

1.3. After this general introduction, we proceed to the algorithm of the study under review. The task is to analyze the *structure* and *texture* of the corpus in question and, on that basis, to identify certain lexical-semantic relations in the specialized language of law. The hypothesis is that the semantic oppositions of the concepts, the relations of *inclusion* and *exclusion* in the lexicon, as well as the relations of *repetition*, *complementation* and *paraphrase* in the syntagm are thematically restricted. In its final version, this hypothesis goes as follows:

---

[2] There are studies of the collocations concerning Slavic languages – cf. Cermak and Holub 1982, Barakova 1995, Reuther 1996. In some works there is no clear distinction between the terms *collocation* and *valence*. Just let us remind that the later is marked as connected with the theory of Tesnière (1953) about verbial grammatical environment.

[3] Needless to say, the tasks formulated by the International Association of Forensic Linguistics are not confined to the "testing" of linguistic models in the legal system of communication, but also deal with, for example, how speech acts function with the application of Schegloff's conversational analysis to court dialogues, etc.

"There cannot be a good lexicon of the whole language!" The lexicon of the entire language is made up of numerous sub-lexicons, which differ by subject and genre – just as coherent texts do not include the entire language but are divided by subject and genre depending on the cultural contexts of sociality.[4]

To prove this hypothesis, the study applied 12 computerized procedures of three types: automatic, manual, and manual-automatic. The 12 procedures are:

1. Identifying the words with the highest frequency, using the NoteTab-Light software. This is an automatic procedure, in which the analysis proceeds from the syntagmatic to the paradigmatic level.

2. Identifying the grammatical characteristics and grammatical classes of words – a manual procedure, since I am not aware of the existence of software that can identify the morphological features of a given corpus in the Bulgarian language and, most importantly, in the Cyrillic alphabet.

3. Identifying – within the grammatical classes – the most frequent words, or the so-called rank list. This procedure is entirely automatic (NoteTab-Light).

4. Identifying the grammatical forms of the words – not in general, but as they occur in the corpus – known as tagging. This procedure is manual-automatic, using NoteTab-Light.

These first four procedures may be defined as *content analysis of grammatical classes*. They pave the way for the subsequent essential analysis of the *texture* (cohesion) of the text.[5]

5. Identifying the collocations (word combinations) of the most frequent nouns, verbs, adjectives, participles and adverbs (ten samples each). Here the analysis proceeds from the paradigmatic to the syntagmatic level because the criterion for the *keyword* in collocations is based on the concept of *frequency* established by means of the first four procedures. The software applied here is FindText.[6]

6. Identifying the repetitions in the text or analyzing tautologies (vertical repetitions of stems, not of forms). This step actually repeats Step 1, but is required here for an analysis not of collocations but of coreferences. The findings are a reference point for the next step: analysis of syntagmatic variability.

---

[4] This framework was outlined even by Malinowski, but can also be found in the works of his followers Halliday and Hasan (1986), as well as in the similar interpretations of van Dijk (1981) and Paul Ricoeur (1986).

[5] In fact, the results of these analyses have been published in Angelov 2000.

[6] This program was prepared especially for the purposes of this study by M. Vojnova.

7. Identifying variability in the cases of same or similar meaning. In fact, this is an analysis of synonymy (which is regarded as including hyponymy and meronymy). The analysis proceeds from the paradigmatic to the syntagmatic level, and the procedure is manual-automatic, using the Presto software for encoding.

8. Analysis of cataphoric and anaphoric coreferences, so-called anaphora resolution. This procedure is applied from the syntagmatic to the paradigmatic levels, using Presto once again – manual-automatic encoding.

9. Identifying homonyms in the text. This is the so-called semantic discrimination or concordance. The procedure is from the syntagmatic to the paradigmatic level, with back testing – verification in the syntagm aimed at isolating syntactic semantizations. The procedure is manual-automatic: homonyms appear in the automatic search, but have to be described and analyzed, i.e. disambiguated.

10. Identifying antonyms – the procedure is automatic for antonyms of the X/not-X type, but manual for antonyms with different roots. The approach is from syntagm to paradigm in the first case, but from the paradigm to syntagm in the second case.

11. Observations on recurrence (repetitions of roots and morphemes; also analysis of syntactic agnation and parallelism). This is, in fact, isolating derivatives – a comparatively simple task for the computer, which can apply FindText to search for and find whole words as well as morphemes.

12. Isolating terms and terminological phrases. This sub-task is a generalization of the previous procedures and mainly of the analyses of collocations and coreferences.

2. Results

The results ought to confirm the principles of linguistic variability formulated by R. Jakobson. As is known, those principles are selection and combination (cf. Waugh and Monville-Burston 1990). The same principles, called *restrictive selection*, are formulated by Katz and Fodor (1963), who regard *selective restriction* (cf. Johnson-Laird 1988) as a projection of the paradigm on the syntagm, i.e. *generation*. Computational corpus analysis is modeled rather on the reverse process of generation, i.e. *reception*; in other words, it involves an approach from syntagm to paradigm or, more precisely, *identification (resolution)* in the syntagm and *semantization (comprehension)* in the paradigm (even though identification, as we also know from phonology, is systemic-structural).

2.1. The most frequent meaningful words

2.1.1. Procedures 1 to 4 identified the following words as the most frequent meaningful nouns: година 'year,' лице 'person/party,' закон 'statute,' съд 'court,' право 'law,' срок 'period,' съвет 'council/board,' средство 'means/vehicle,' ред 'procedure,' свобода 'liberty'; in the course of the analysis of the grammatical forms, however, their order changed: закон 'statute,' лице 'person/party,' право 'law,' ред 'procedure,' съд 'court,' свобода 'liberty,' срок 'period,' съвет 'council/board,' средство 'means/vehicle,' сила 'force.'

2.1.2. The most frequent verbs in legal discourse (excluding the verb съм 'to be' in all its forms, as well as мога 'can' – and the forms може/не може 'can/cannot' respectively, and има/няма 'to have/have not') are: наказвам 'to punish,' извършвам 'to perform,' определям 'to determine,' издавам 'to issue,' отговарям 'to respond/be liable/be in charge,' налагам 'to impose,' допускам 'to admit,' произнасям се 'to pronounce,' упражнявам 'to exercise,' стигам 'to achieve,' провеждам 'to conduct.'

2.1.3. The most frequent adjectives in legal discourse are: министерски 'ministerial,' български 'Bulgarian,' народен 'national,' превозен 'shipping,' административен 'administrative,' вътрешен 'internal,' трудов 'industrial,' държавен 'state,' професионален 'professional,' предходен 'preceding.'

2.1.4. The most frequent participles (which are very important because they indicate universal roles of the persons) are: осъдения 'the convict,' разпитвания 'the interrogatee,' заподозрения 'the suspect,' осигурения 'the insured party,' спечелилия 'the successful party,' застраховащия 'the insuring party,' превозващ 'a carrier,' следващия 'the next,' спасяващия 'the rescuing party,' отсъстващия 'the absent,' наказвания 'the punished,' запретения 'the legally incapable,' определен 'specified.'

2.2. Collocations of 10 of the most frequent meaningful nouns

2.2.1. The noun collocations show that the word закон 'statute' occurs in the following set phrases: ред, определен със закон ... 'procedure regulated by statute ...,' случаите, предвидени в закона ... 'cases provided for in the statute ...,' лицата, установени от този закон 'the persons specified by this statute' The collocation по силата на закона 'by force of statute' is also a set phrase, along with по ред установен в закона ... 'according to a procedure established by statute ...,' според този закон 'according to this statute.' In addition, the most frequent verbal collocations of the word закон 'statute' are the following:

прилага 'applies,' урежда 'regulates,' възлага 'vests,' влиза в сила 'enters into force.' This analysis indicates that the noun закон 'statute' has a special, almost "sacred" function in legal texts. Through it the text identifies itself as legal, and this word is thus at the top of the systematic hierarchy, collocating with a whole range of adjectives that specify its meaning.

2.2.2. The word лице 'person/party,' which is the abstract legal subject that becomes an обект на третиране от закона 'object of the law,' has a similar function. This word occurs in the following collocations: пътник е лице, което ... 'a passenger is a person who ...,' водач е лице, което ... 'a vehicle operator is a person who ...'; заето лице е всяко лице, което ... 'an employee is any person who ...,' безработно лице е всяко лице, което ... 'an unemployed (person) is any person who ...' Hence, this type of definitions in statutory texts plays the role of *social casting*. Apart from social casting, however, we also have *court casting*: вещо лице 'expert witness,' трето лице 'third party,' осъдено лице 'convicted person,' уличено лице 'guilty party,' укривано лице 'harboured fugitive/absconder,' задържано лице 'apprehended person/person in custody/detainee,' регистрирано лице 'registered person,' призовано лице 'subpoenaed/vouchee,' подставено лице 'straw man/dummy,' пострадало лице 'wronged/injured person,' длъжностно лице 'official/office holder,' etc.

2.2.3. The word съд 'court' also plays the role of a supreme institution and main agent which realizes the law. In addition, the court-qualifiers indicate the court hierarchy: Конституционен съд 'Constitutional Court,' върховен съд 'Supreme Court,' върховен касационен съд 'Supreme Court of Cassation,' окръжен съд 'district court,' районен съд 'regional court,' апелативен съд 'appellate court,' Европейски съд 'European Court,' etc. The most frequent verbal collocations are: разглежда се от съда 'heard by the court,' определя се от съда 'adjudicated by the court,' прекратява се от съда 'dismissed by the court,' etc. For its part, съд 'the court' разрешава 'grants,' назначава 'appoints,' поканва 'invites,' изслушва 'hears,' постановява 'rules.'

2.2.4. The word право 'right/law' is polysemous. In the phrases всеки има право да търси ... 'everyone has the right to seek ...,' всеки има право да информира ... 'everyone has the right to inform ...,' служителите имат право да се сдружават ... 'employees have the right to associate ...,' this word collocates with the verb има 'to have' and, respectively, няма 'to have not,' and it is clear that its meaning is close to that of the word свобода 'liberty' or the word разрешение 'permission.' Beyond the collocations with има/няма 'to have/have not' and if it is modified by a qualifier, право has a different meaning – the legal metatext and body of legal standards, i.e. 'law': международно

право 'international law,' наказателно право 'criminal law,' гражданско право 'civil law.' Here the meaning of the word право is close to that of the word закон and likewise indicates a classification and hierarchy of the body of legal texts.

2.2.5. The word срок 'period' occurs in the following set phrases: седемдневен срок от получаване на призовката 'within a period of seven days after receipt of the summons,' изтичане на срока 'expiration of the period,' срокът започва да тече от ... 'the period begins to run from ...,' установени срокове 'established periods,' давностни срокове 'limitation periods,' etc. The analysis shows that this word collocates with variable numerals: срок от 5 години 'a period of five years,' срок от 3 години 'a period of three years,' etc.

2.2.6. The word съвет 'council/board/advice' requires a modifying adjective that specifies its meaning in 90% of the cases: in statutory texts this word is used not in the sense of 'advice' but of 'council/board,' i.e. 'a group of people' with particular functions: Министерски съвет 'Council of Ministers,' Надзорен съвет 'Supervisory Board,' Управителен съвет 'Management Board,' Корабен съвет 'ship's crew council,' Съвет на директорите 'Board of Directors,' etc.

2.2.7. The word средство 'means,' is as abstract as the word лице 'person/party.' It can have countless referents and requires qualification. The frequency of this word, however, is highest in the Road Traffic Act, where it almost always refers to means of transport: моторно превозно средство 'motor vehicle' or пътно превозно средство 'road vehicle.'

2.2.8. The word ред 'procedure' constitutes – just as the words закон and право (in the sense of 'law') – the statutory texts, indicating the algorithm (sequence) of legal proceedings: общия ред 'the standard procedure,' по исков ред 'by an action proceeding,' по съдебен ред 'judicially/through the courts,' etc.

2.2.9. The noun свобода 'liberty' occurs in the phrase лишаване от свобода 'deprivation of liberty' in 99% of the cases. This is in fact a euphemistic usage of the word затвор '[term of] imprisonment' or наказание-затвор 'penitentiary punishment.'

2.2.10. The word сила 'force' occurs in the phrases: законът влиза в сила 'the statute enters into force,' решението влиза в сила 'the decision enters into force.' That is because in the pragmatics of legal texts there are two important initiations of the text: first, upon its enforcement in regard to the whole society, when it can be read by everybody and has a preventive function; and second,

when it is read in court (reading aloud) – when the law *is realized* or, in legal terms, *is applied* to the relevant persons.

3. Conclusions

As the review of nouns shows, statutory texts *prescribe* specific collocations of the keywords, i.e. we have formulas and set phrases. These formulas, however, are also indicative of the specific pragmatics of this type of texts, which are obviously "designed" to cover, place within a certain framework and regulate the whole of society. In practice, the laws realize important social distances, which result from the social and textual interaction; they divide society in two – *law-abiding citizens* and *law-breakers,* while assigning a special role to jurists, who are *intermediaries* and *agents* (surgeons, if you will) in this type of interaction. As we have seen, however, this is sometimes hidden by euphemisms – лишаване от свобода 'deprivation of liberty' means затвор 'imprisonment,' whereas платежни средства 'means of payment' equals пари 'money.'

3.1. For considerations of space, I will not discuss typical collocations of the most frequent verbs, adjectives and adverbs. As regards the verbs, I will only note that they are very few in number – the verbal frequency in statutory texts is only 6.23%, versus 18.92% in fictional and 17.2% in spoken conversational texts. Besides, most of the verbs do not have autonomous concrete meaning (such verbs may refer to different kinds of action): извършвам 'to perform,' провеждам 'to conduct,' упражнавам 'to exercise.' This, however, does not apply to the verb наказвам 'to punish,' which is very frequent – but mainly in criminal law, and is used most frequently in the reflexive impersonal construction се наказва 'shall be punishable.' The doer is thus an unspecified, impersonal subject (although everybody knows that this is закон 'the law' and в името на закона 'the court in the name of the law'), but those impersonal phrases show that in fact there is an *avoidance of responsibility,* as if the court *was apologizing* for its adjudicative power. The court, as supreme implementer of the law, also hides behind other impersonal constructions such as: се допуска 'it shall be assumed,' се налага 'it shall be necessary,' се определя 'it shall be determined,' се счита 'it shall be considered,' etc. In addition, the verbs in these texts are entirely in the third person singular and plural, and in the present tense. Perfective verb forms which, as is known, collocate with modal verbs or future or past perfect forms, are avoided as often as possible. The modal forms съдът може да постанови 'the court may hold,' съдът може да наложи 'the court may impose' are exceptions.

3.2. All those conclusions are drawn on the basis of the analysis of syntagmatic structures and, moreover, of the analysis of collocations only. The analysis of coreferences requires special attention. As regards paradigmatic structures such as homonymy, synonymy and antonymy, I will only note that even though synonymy and homonymy are deliberately avoided in legal texts, both types of semantic variability occur in these texts. As is known, homonymy is easily eliminated in syntagmatic propositions, unless it results from some deeper grammatical or pragmatic ambiguity. In NLP terms, this elimination is called disambiguation. As a method of analysis (i.e. a form of reception), disambiguation is related to the transition from syntagmatic to paradigmatic. Conversely, in text generation, paradigmatic homonyms are referred to and tested in propositional string disambiguation.[7] In other words, the polysemy of expressive devices is resolved by lengthening the propositional string. Redundancy may thus be tautological, but is in most cases synonymous and designed to help the listener (reader) in coping with homonymy. In some cases this might even be an elaborate precaution against misunderstanding. The accumulation of synonyms thus has a cumulative semantic effect and, from the perspective of theme-rheme relations, the rhemes in such synonymous propositions are coreferent only, and not referent.

3.3. Finally, I would like to draw attention to the phenomenon of *flexive disambiguation*, or avoiding ambiguities through suffixation. Let us take a random example: the word дърво 'tree/wood/timber.' To check the meaning of this word, we use the plural forms: дърво-дърва 'wood/wood'[8] and дърво-дърветa 'tree/trees.' This disambiguates the polysemy and also changes the references. Flexion thus plays the role of *microlocation* and of *disambiguator* of lexical polysemy. This also applies to words which have two meanings in the singular and just one in the plural. Thus for example, the word право 'law' is a singularia tantum in one of its two meanings: 'a body of legal documents,' 'meta-language of the law.' In its second meaning, the word право 'right' has a plural form and is synonymous with the word 'liberty' in the sense of 'rights and liberties.' This is how the microlocation of the suffix is a factor for the disambiguation of homonymy.

But the semantic network of lexical synonyms and antonyms also plays the role of a system corrector even without a syntagmatic test. We can draw the

---

[7] We should note here that computers do not make a strict distinction between homonymy and polysemy.

[8] This example is hard to be translated, but even in English we have **wood** (singulare tantum) – hard solid substance of the tree below the bark; and **woods** (plurale tantum) – area of land covered with growing trees (not so extensive as forest).

conclusion that semantic relations at the paradigmatic level have ordered referents in the mental lexicon, but they need to be specified in the syntagm, and occasionally acquire secondary semantizations. Grammatical devices can serve as a corrector in some cases only (the article morpheme can also be such a corrector). In most cases, this role is played by lexical collocations.

In conclusion, I would like to note that flexive disambiguation requires special attention from the perspective of computational lexicography, and offers yet another argument for approaching mental system reality as adequate to both sensory and semiotic reality.

Literature

Angelov, A. 2000. „Лингвистические измерения современного болгарского юридического дискурса" (Linguistic Dimensions of the Bulgarian Legal Discourse), L. Zybatow (ed.), *Sprachwandel in der Slavia*, Frankfurt am Main: Peter Lang – Linguistik International. T. 1, 423-448.

Barakova, P. 1995. Баракова, П. Прилагателни с минимален колокационен обхват в съвременния български книжовен език. *Български език*, 1995/3.

Čermák, F., Holub, J. 1982. *Syntagmatika a paradigmatika českého slova. Valence a kolokabilita*. Praha: Státní pedagogické nakladatelství.

Cruse, D.A. 1986. *Lexical Semantics*, New York: Cambridge University Press.

de Beaugrande, R.-A., Dressler, W.U. 1981. *Introduction to Text Linguistics*. London: Longman. Translated in Bulgarian: Р. дьо Богранд, В.У. Дресслер, С. Стоянова-Йовчева, 1995. *Увод в текстовата лингвистика*. София: Университетско издателство "Св. Климент Охридски".

Firth, J.R. 1957. „Models of Meaning", *Papers in Linguistics 1934–1951*. London, 190-215.

Halliday, M.A.K. 1966. „Lexis as a Linguistic Level", C.E. Bazell, J.C. Catford, M.A.K. Halliday, H.R. Rolins (eds.), *In memory of J. R. Firth*, London: Longman, 148-162.

Halliday, M.A.K., Hasan, R. 1976. *Cohesion in English*, Longman.

Halliday, M.A.K., Hasan, R. 1986. *Language, context and text: aspects of language in a social-semiotic perspective*, Oxford University Press.

Hasan, R. 1984. „Coherence and Cohesive Harmony", J. Flood (ed.), *Understanding Reading Comprehension*, Newark, DE: International Reading Association, 181-219.

Johnson-Laird, P.N. 1988. „La représentation mentale de la signification", *La science cognitive*. Vol. 115, 53-69.

Katz, J.J., Fodor, J.A. 1963. „The structure of a semantic theory", *Language*. 39, 170-210.

Kniffka, H. (ed.) 1990. *Texte zu Theorie und Praxis forensischer Linguistik*, Tübingen: Max Niemeyer Verlag.

Lyons, J. 1977. *Semantics*. Vol. 1-2, London: Cambridge University Press.

Miller, G.A., Fellbaum, Ch. 1992. „Semantic Networks of English", B. Levin, St. Pinker (eds.), *Lexical & Conceptual Semantics*, Blackwell, 197-230.

Reuther, T. 1996. On Dictionary Entries for Support Verbs: The Cases of Russian VESTI, PROVODIT' and PROIZVODIT'. In: Leo Wanner (ed.) 1996. *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam: John Benjamins, 181-208.

Ricoeur, P. 1986. *Du texte al'action. Essais d'herméneutique II*. Editions de Seuil. Translated in Bulgarian: П. Рикьор 2000. *От текста към действието. Херменевтични опити. Т. 2*, Съставителсвто – Иванка Райнова, София: Наука и изкуство.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tesnière, L. 1953. *Esquisse d'une syntaxe structurelle*, Paris.

van Dijk, T.A. 1981. *Studies in the Pragmatics of Discourse*, The Hague: Mouton. Translation in Russian: ван Дейк, Т.А. 1989. *Язык, познание, коммуникация*, Москва: Прогресс.

Waugh, L.R., Monville-Burston, M. (eds.) 1990. *On Language/ Roman Jakobson*. Cambridge, Mass.: Harvard University Press.

Wierzbicka, A. 1999. Вежбицкая, А. *Семантические универсалии и описание языков*, Москва: Языки русской культуры.