Danko Šipka
University of Munich, Germany

CDU 801.3

USAGE LABELS NETWORK: AN APPROACH TO LEXICAL VARIATION[1]

1 *State of the art*

1.1 The Problem of lexical variation is frequently addressed within the linguistic community. Its complexity and the broad implications of any possible solution have considerable appeal among theoretical linguists. Lexicographers, in their turn, have been forced to address it in order to provide dictionary usage information, which is normally done by means of dictionary labels such as: *American English, obsolete, slang*, etc. An insightful overview of the relevant lexicological approaches, as well as some lexicographis projects is provided in Lipka (1990). The most exhaustive sociolinguistic classification, however, can be found in Preston (1986). Lexicographis treatments of lexical variation have been addressed in numerous papers listed in Zgusta (1988).

1.2 A careful review of the papers mentioned above as well as my investigation of several Slavic, German, and English dictionaries (described in Šipka 1992 in print), allows us to formulate the following general remarks about the problem:

a. the underlying criteria for the categories distinguished are, in most cases, neither clearly stated nor recognizable,

b. the same holds for the hierarchization of the categories,

c. frequently, the different-level categories are treated as if they were same-level categories,

d. there is no common agreement about the underlying criteria or about the categories and their hierarchization.

1.3 All this can be exemplified by means of the label categories distinguished in several prestigious slavic (mostly academic) dictionaries. Their non-consistency can be

[1] I would like to express my gratitude both to the Alexander von Humboldt Foundation, which made my research on the topic possible, and Prof. Leonhard Lipka, of the University of Munich, my adviser during the Humboldt Post-Doctoral Program, for his support and various useful comments on former versions of the manuscript. I am furthermore very grateful to Dr. Hans-Jörg Schmid for his useful comments of the manuscript.

observed in comparison with a consistent list of categories, like the one presented in Lipka (1990: 23):

(a) region	
(b) social group	user
(c) *field of discourse*	
(d) medium	language use
(e) attitude	

The label categories used in Slavic dictionaries, given here in literal translation, are stated as follows:

Russian dictionary (ANSSSR): transferred meaning, jocular-ironical, non-literary words, terms (1950: I, XII)

Bulgarian dictionary (BAN): non-literary, functional style, historical, emotional-expressive, frequency, style character change (1977: I, 25)

Slovenian dictionary (SSKJ): semantic, terminologic, style, expressive, temporal-frequency, normative (1979: I, XX)

Polish dictionary (PAN): geographic, thematic, chronologic, expressive (1958: I, XXXIX)

Slovak dictionary (SAV): non-literary, style, emotional-expressive, temporal, normative (1959: I, XI-XIII)

Serbo-Croatian dictionary (MS): professional terms, archaisms, neologisms, vulgar and slang words, hapax legomena (1967: I, 11)

Macedonian dictionary (IMJ): style markers (mentioned only in general sense) (1961: XII).

All the dictionaries mentioned are similar to a great extent in their approach. The differences, therefore, are not caused by the dictionary type, the intentions of the compilers or the needs of potential users.

A similar situation can be observed with German (DUDEN 1967ff), as well as English dictionaries (LCED 1985, Collins 1986). Furthermore, these findings are supported by several metalexicographic papers (eg. Ludwig 1982, Schippan 1987).

1.4 In order to overcome the situation stated above, we propose the construction of a "usage labels network" and its algorithm for handling ambiguity and synonymy.

2 *Starting assumptions*

2.0 The basis for the network and its algorithm are some relatively simple facts of human perceptive and creative abilities. These facts can be roughly described as follows.

2.1 When reading or writing a text, one normally knows what is the object of the reading or writing, i.e. one is aware of the "text type" in question. Usually, we know whether we read a newspaper sport section, a government document, a poem, etc. Similarly, we are aware that we are producing a letter to a friend, a novel, an official statement, etc.

In such situations we always have in mind which lexemes are allowed in certain contexts. Consequently, we expect to read or to use not all the lexical units, but only the ones that are justified by the context. Thus, when we meet a form that can belong to two or more different meanings, or when choosing the most appropriate synonym, in most of the cases the determining factor will be the type of context.

To illustrate this, let us use two simple examples.

When reading a newspaper report on a chess tournament and encountering in it the word *partia*, a speaker of Russian assumes that the meaning is 'game' since the context disfavors the other meaning, 'political party'. When reading a Communist Party document, our imaginary Russian speaker is in the opposite situation, i.e. he expects 'political party', not 'game'. This means that the type of context (ie. not context itself) operates as the disambiguator in this particular case.

When producing a vulgar joke, or an official statement, a speaker of English may use the phrases *to kick the bucket* or *to pass away*, respectively. In neither of the situations will one hesitate about the decision. It would be incorrect to use the phrase *to kick the bucket* in an official statement, say in a newspaper; and it is ridiculous to use the phrase *to pass away* in unofficial communication, unless one is trying to be ironical.

2.2 It follows that before reading or writing, one has eliminated all the senses and the forms (ie. the 'lexical units' in Lipka's sense) that do not apply in the particular situation. In all such cases, therefore, it would be a pure waste of time and energy to search for contextual clues, and, quite obviously, one does not do this.

Basically, what we have in our mind when reading or writing a text can be roughly described as the 'labels' (in the sense defined here), both for the texts and lexemes, as well as the rules which determine the relations between, as well as the rules which determine the relations between the text and the lebels for the lexemes, or the 'lexical units' of a 'lexeme'. Or, in other words, both texts and lexemes are classified in pigeonholes, so that some senses of a lexeme (ie. 'lexical units') fit into a certain text's pigeonhole, and the others do not.

Briefly, prior to reading or wrtiting we must have in our mind:

1. the "usage labels network":
    a. labeled text,
    b. labeled lexemes or 'lexical units' ie. different senses of lexemes,
    c. the rules which govern the text vs. lexeme label relations,
2. the label of the text we are reading or writing.

For example, when handling the Serbo-Croatian pair of 'lexical units' that fall together in one lexeme *čast*, namely 'honor' and 'part', the first meaning will be labeled as contemporary, the second as obsolete. The text of a newspaper from the year 1989 will be labeled as contemporary. There is a rule that the text label 'contemporary' significantly reduces the probability of occurrence in the text of any 'lexical units' labeled as 'obsolete'. Consequently, a reader does not expect the second meaning of the homonymous or polysemous lexeme(s) *čast*, and a writer is not going to use such a meaning.

The same applies to synonyms. If this were not so, we would accept, for example in The New York Times, the sentence: *The president kicked the bucket this morning at 6:30* as quite normal.

2.3 For the reasons explained above we can assume that these labels and rules are an inherent part of language competence, not just the facts of language performance. It is, therefore, quite legitimate to try to establish a systematic list of labels and to describe the rules which govern their usage.

3 *Construction*

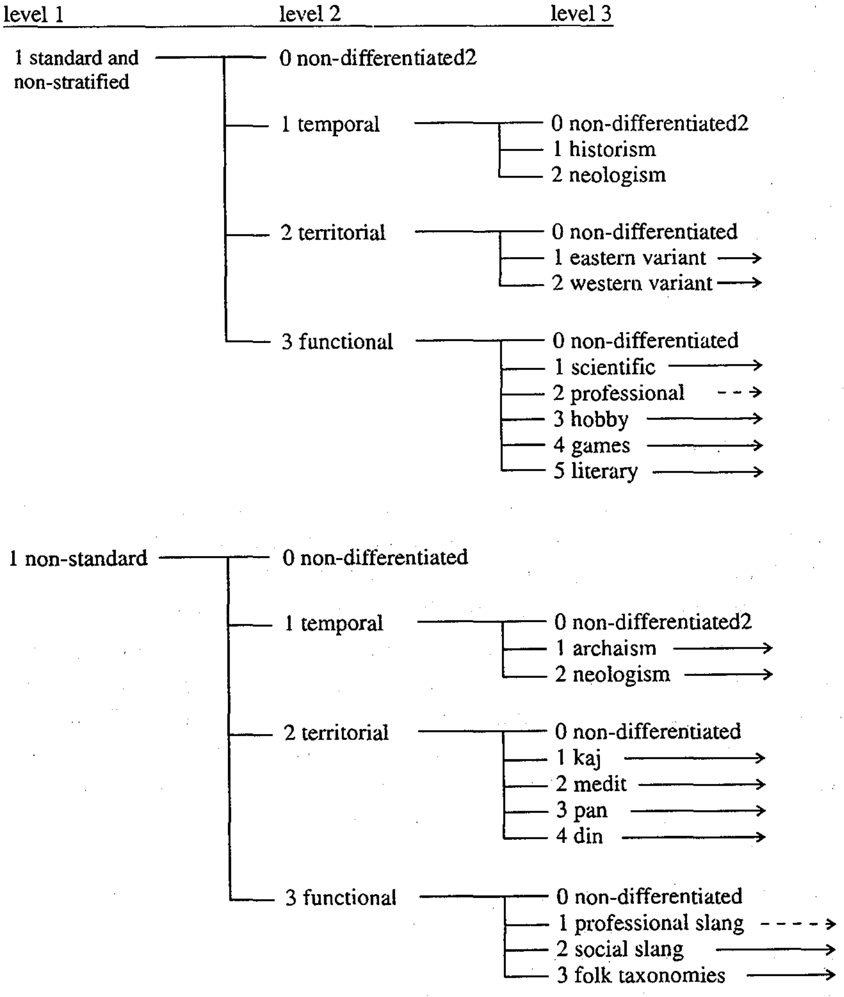
3.0 The first distinction is to be made between those usage labels which cannot be stated as the probability of a lexeme or 'lexical unit' to be found in a particular text group, and the ones which can be treated that way. Note that we have in mind only usage labels, not, for example etymological ones.

The former comprise the following categories, with the respective values given in brackets: **frequency** (frequent, usual, rare, individual...), **expressive** (derogatory, vulgar, jocular, familiar...), **personal** (baby talk, masculine, feminine), **referential** (used of ...) labels. These labels are not directly useful for our purposes, due to the fact that they can not be formulated as the lexeme-text relation. Cf. the distinction between "dictionary labels" and connotations in Lipka (1990: 14-26, 63-67).

The latter, however, function as the basic elements of the following usage labels network.

3.1 Hierarchy is the governing principle of the network. Sometimes, for example in Miller et al.'s (1990) database *WordNet*, this is called a "lexical inheritance system". Herarchic structure in the lexicon is particularly relevant for nouns, while for adjectives antonymy is more important. Within the framework of 'sense-relations' developed by John Lyons (cf. Lipka 1990: 140ff), hierarchical relations in the lexicon are described with the help of the concept of 'hyponymy'. In the work of anthropologists and cognitive linguists, 'natural (or folk) taxonomies' were distinguished from 'scientific (or technical) taxonomies' (cf. Lipka 1990: 155f). The proposed network consists of usage labels mutually related on the basis of subordination and coordination. It is a tree-like structure with top-down subordination, and terminal points at each level. Thus, if the difference between two lexemes, or between a text and a lexeme is obvious at the highest level, there is no need to go further.

A draft of the highest levels for Serbo-Croatian can be presented as in the diagram 1:



[2] Cf. the discussion of "markedness" in Lipka (1990: 63ff).

The arrow indicates further differentiation, while 0 indicates the terminal point of a branch.

Thus, for example the category 'games/plays', covered in Serbo-Croation by one word (*igre*) and, respectively, one concept *igre* branches further as in the diagram 2

(2):

4 games/plays

- 1 dancing
  - 1 dance
  - 2 folk dance
- 2group open-air games
- 3 table games
  - 1 board games
  - 2 card games
- 4 sports
  - 01 bal games
  - 02 gymnastics
  - 03 track and field
  - 04 hiking and alpinism
  - 05 cycling
  - 06 auto-moto
  - 07 fighting
  - 08 winter
  - 09 water
  - 10 riding
  - 11 aero
  - 12 hunting and fishing

This example is also interesting to show that there does not have to be a single common denominator for all members of a category. The German philosopher L. Wittgenstein used the very same example to exemplify his concept of 'family resemblance' (cf. the discussion of his approach in Aitchison 1987: 74ff).

3.2.0 The network is based both on ontological (or ontognoseological) and linguistic grounds. The main sources of the ideas for the network generation were lexicologic, socio-, and psycholinguistic handbooks; general, specialized and frequency dictionaries; numerous papers (lexicological on vocabulary stratification, metalexicographical on dictionary labels, papers form computational linguistics on lexical data basis); and finally, library classificatory systems: decimal classification, Dewey, etc. (cf. review of the systems in Bakewell 1978).

3.2.1 The initial binary branching is the split between standard and non-differentiated versus non-standard. The markedness of the non-standard group reflects its higher peculiarity, when compared with the standard one. This first branching shows already one great advantage of the network, i.e. if the a text is labeled as standard, and a lexeme or a 'lexical unit' as non-standard, the determination of the possibility for a lexeme to be in that text is already accomplished.

3.2.2 Further sub-branching in both categories is a threeslot split (temporal, territorial, and functional), with separate branchings for each of them. While the first branching was obligatory, all the others are facultative. Most of the prepositions and conjunctions in all languages, for example, are labeled only as standard/non-differentiated, with no further differentiation at all.

3.2.2.1 The temporal slot in both categories comprises two binary branchings. A word can be unmarked or marked, and the marked ones can be **neologisms** or **historisms** (in standard and non-differentiated groups), i.e. **neologisms** or **archaisms** (in non-standard group). The notion of neologism is rather indistinct. The problem of their determination is where to set the temporal limit after which a word can be considered a neologism. The difference between standard and non-standard neologisms is solely the one between their superordinated categories. Historisms and archaisms, however, are distinct in yet another manner. The lexemes that are used only about denotata which do not exist any more are considered historisms (such as *bey* ie. 'a title in the Turkish Empire', *knight*, etc.). They are used in standard language, but only in connection with these denotata. Archaisms, on the other hand, became obsolete for their form, and their denotatum is referred to in standard language by another word (eg. *albeit*).

3.2.2.2 The territorial slot is a language-specific category. The slot presented above pertains to Serbo-Croatian. In the standard category it corresponds to national variants: **serbian (eastern)** and **croatian (western)**. Non-standard territorial branching only partially corresponds to Serbo-Croatian dialects due to the fact that various dialects share a very similar cultural background, and thus similar lexical influences, which in its turn leads to identical lexical strata. For example, the lexemes labeled as **mediterranean** can be found both in the ča- and što- dialect, which are, in dialectologic perspective, two quite different entities.

3.2.2.3 The functional slot is certainly the most intricate one. The differentiation in both standard and non-standard categories should be universal, at least in an European or Europe-based culture, ie. a culture with an "European" set of values and way of thinking. In the standard category, it follows the differentiation of human activities. Therefore, we have the categories corresponding to **scientific, professional, leisure (hobby** and **games)**, and finally **literary** activities. Non-standard differentiation, on the groups, i.e. members of certain professions **(professional slang)**, social strata **(social slang)**, or very broad, uneducated, mostly rural population strata **(folk taxonomies)**.

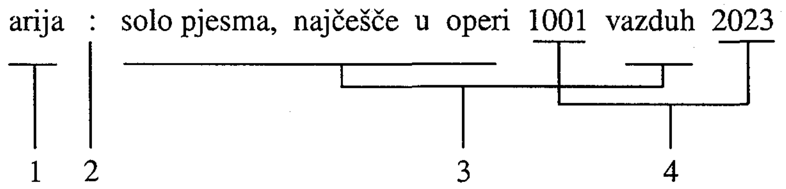
3.3 Although one might find some other approach to the categories more useful (for example to have only "professional" and "leisure" instead of our five standard functional categories), the basic principles of the network, primarily the very idea of tree-structureness with the topdown flow, should substantially contribute to our understanding of lexical variation.

Another question, not to be discussed here in detail, is now to assign the labels to the lexemes. One can use non-gradual labels (eg. *slang*) and, therefore, label only lexemes that clearly belong to the category in question, which was our approach in testing the network. It is, however, also possible to set a scale, for example between 1 (clear member of a category) and 0 (clear non-member), and thus have gradual labels (such as *slang 0.1, slang 0.7, slang 0.8,* etc.), following thus the idea of "gradeience" very much in current usage in linguistics.

4 *Testing*

4.1 The network has been tested on a database consisting of 1105 Serbo-Croatian homonymic nests, containing a total of 2287 labeled lexemes. The database entries were constructed as in the diagram 3

(3):



1 - form shared by the homonyms
2 - separator
3 - meanings (*aria* vs. *air*)
4 - label codes.

The label codes follow the numbers on the graph presenting the highest levels of the network. So the first place show if it is a standard (1) or non-standard lexeme (2), and the rest of them temporal, territorial and functional differentiation. Thus, in our example: 1001 = standard temporally and territorially non-differentiated lexeme belonging to scientific terminology vs. 2023 = non-standard temporally non-differentiated Mediterranean folk-taxonomy. The database was planned to be reusable: it is used to test the network, as the basis for disambiguating software, and finally, to analyze several categories of variation in Serbo-Croatian.

The nests have been derived from the 6-volume dictionary by Matica Srpska (1967-76). Their labelling was based on various dictionary labels, as well as my own native speaker competence. Only the three initial levels of the network were applied. There were possible results: the nests could be:

a. solved with just the first three levels,
b. solvable with further levels,
c. unsolvable.

The network has been tested on the example of homonyms since the idea was that if lexical differentiation is so clearly stated that it can be efficiently used for disambiguation, then it is plausible to expect that it will function in all other casses as well.

4.2 As the final result we had 719 (or some 70%) nests solved with the first three levels, 342 nests were solvable with further levels, and finally 98 unsolvable nests remained. This proves the network to be highly efficient: as more than 90% of homonymic nests can thus be solved.

More-than-two-member nests having some binary relations solved, while other further solvable, or unsolvable were counted in two or three groups. If, for example, we have a nest ABC consisting of the binary relations AB, AC, BC, and AB being solved, AC solvable, BC unsolvable, then all possible results are counted for that nest. This is why the final result is 1159 (719 + 342 + 98), and there are only 1105 nests.

Furthermore, it is interesting that 1565 lexemes (71,69%) were, one way or another, stratified, while only 618 (28,31%) were marked only as standard or non-stratified. This shows that the status of a wide range of lexemes is determinable by the network.

5 *Applications*

One can think of numerous applications of the network and its advantages when compared with existing models and practice. The network can be applied to a variety of linguistic and other activities, two immediate ones being lexicography and computational linguistics.

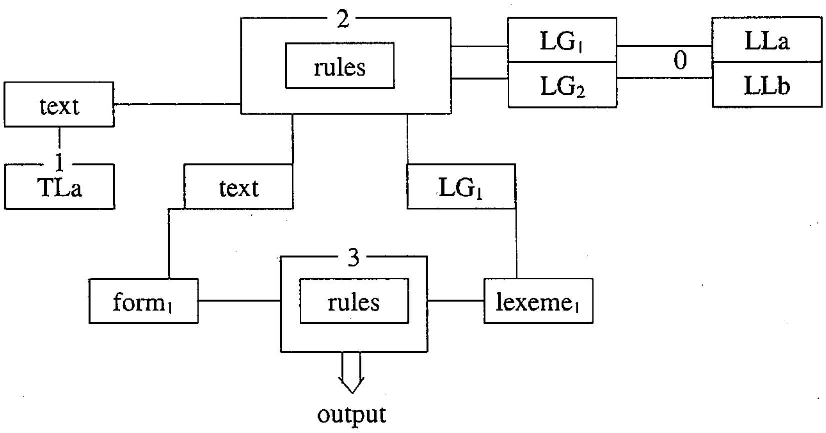
In the field of lexicography, the usage labels network brings a more consistent theoretical approach: both underlying principles and categories can be clearly stated as well as hierarchized. Moreover, there are numerous practical advantages. The lexicographer is offered a solid basis for labelling, the user knows which categories he/she is going to deal with. The network is, furthermore, adjustable to the dictionary type. Thus, a general descriptive dictionary equally develops all the branches with dept of the branching depending on its volume, intentions, user needs, etc. A specialized dictionary, on the other hand, in detail sub-classifies only the branch representing its field, while the others remain only roughly differentiated (eg. using only the first three levels of the network).

There are two main applications of the network in the field of computational linguistics, both being based on matching lexeme versus text labels. The first application concerns the choice of synonyms. Most commercial text-processors offer as their thesaurus a list of synonyms without usage labels. Applying the usage labels network, this text-processors' option can be substantially improved. The user would choose the text labels (once for each text), so that the thesaurus option offers only the synonyms associated with the lexeme labels that match those chosen for the text in question. Thus, the phrase *to kick the bucket* will be eliminated in a formal text, and so will *to pass away* in an informal one.

The second application within computational linguistics is the one in the process of disambiguation. Present disambiguating procedures are normally based on contextual clues. As it can be seen in numerous papers on the topic (recent surveys can be found in Small et al. 1988, as well as in Batori et al. 1989), this is a troublesome, time-consuming as well as memory-demanding task. In case the network and the matching of the lexeme versus textual labels is applied prior to searching for contextual clues, numerous instances of ambiguity can be solved without long and complicated procedures. The quantity of the cases solvable by means of the network is indicated by the test results tated in 4.2. Of course, there are also non-serious texts, where no disambiguation is intended. In such cases, the network does not disambiguate, it simply reveals the mechanism of a joke, or other non-serious text.

The algorithm for the application of the network in computational linguistics can be stated as in the diagram 4:

(4)



step 0 LG_1/LLa/, LG^2 /LLb/

This is predetermined. We have linked two different labels to the two separate lexicon groups. For example, 'a' means 'contemporary', 'b' means 'obsolete'.

step 1 text/TLa/

A label is assigned to the text. Here too we can imagine that 'a' means 'contemporary'.

Step 2 text/TLa/ <– G1/LLa/
|– LL_2/LLb/

Predefined rules state that a text labeled as 'a' tolerates only the lexemes labeled as 'a'. Thus, only the lexeme group 'one' is passed through. For example none of the obsolete lexemes are allowed in a contemporary text.

step 3 form1(text) lexeme(LG_1)

Predefined rules determine that a form met in the text belongs to a lexeme from lexeme group one.

LG	– lexeme group
TL	– text label
LL	– lexeme lebel
a<b	– a belongs to b
<–	– allows
|–	– blocks
a/b/	– b is assigned to a
a(b)	– a is from the set b

Further applications could be imagined in the fields of language planning (where the network can show which lexemes should be prescribed or suggested in a certain text), terminology (where only the optimal terms are to be selected, using the network), language training (where the network can indicate which lexemes are crucial for, e.g. a doctor, an engineer etc. who is learning a language), artificial intelligence (where the network could support a more effective semantic interpretation), human and machine translation (where the network helps finding optimal translating equivalent), sociology (where the network might reveal attitudes of a group, i.e. shows whether obsolete, impolite, etc. lexemes have been used frequently or not), political science (the network can help a politician to choose the most appropriate lexicon according to the text type, and thus make his speech more effective), law (the network could show the optimal way for stating a legal rule), etc. It is evident, then, that most of the applications are to be performed by the algorithm.

6 *Bibliography*

6.1 *Dictionaries*

ANSSSR. *Slovar' sovremennogo russkogo literaturnogo jazyka.* Moskva - Lenjingrad, 1950-1965.

BAN. *Rečnik na b'lgarski ezik.* Sofija, 1977ff.

ČAVU. *Príruční slovník jazyka českého.* Praha, 1935-1957.

Collins. *Collins Dictionary of the English Language.* Longon and Glasgow, 1986.

DUDEN. *Das grosse Wörterbuch der deutschen Sprache in sechs Bänden.* Bibliographisches Institut Mannheim/Wien/Zürich, 1976-1987.

IMJ. *Rečnik na makedonskiot jazik.* Skopje, 1961-1966.

LCED. *Longman Concise English Dictionary.* Harlow, 1985.

MS. *Rečnik srpskohrvatskoga književnog jezika.* Novi Sad (- Zagreb), 1967-1976.

PAN. *Słownik jezyka poslkiego.* Warszawa, 1958-1965.

SAV. *Slovník slovenského jazyka.* Bratislava, 1959-1968.

SSKJ. *Slovar slovenskega knjižnega jezika.* Ljubljana, 1970-1991.

6.2 *Other references:*

Agricola *et al.* (Hrsg.). 1982. *Wortschatzforschung heute.* Leipzig: VEB.

Aitchison, Jean. 1987. *Words in the Mind. An introduction to the mental lexicon.* Oxford: Basil Blackwell.

Bakewell, K.G.B. 1978. *Classification and Indexing Practice.* London: Clive Bingley.

Batori, Istvan S, Winfried Lenders, Wolfgang Putschke. 1989. *Computational Linguistics. Komputerlinguistik.* An International Handbook on Computer Oriented Language Research and Applications. Berlin: Walter de Gruyter.

Lipka, Leonhard. 1990. *An Outline of English Lexicology. Lexical Structure, Word Semantics, and Word-Formation.* Tübingen: Max Niemeyer Verlag.

Ludwig, Klaus-Dieter. 1982. "Zur normativen, konnotativen und stilistischen Angaben in Wörterbucheintragungen." in: Agricola 1982.

Miller, G.A. *et al.* 1990. "Introduction to WordNet: An Online Lexical Database", *International Journal of Lexicography* 3, 235-301.

Preston, Denis. 1986. "Fifty Some-Odd Categories of Lexical Variation", *International Journal of the Sociology of Language* 57, 9-47.

Schippan, Tea. 1987. "Zum Charakter "stilistischer" Markierung in Wörterbuch" in: Klaus Welke und Renate Neurath (Hrsg.). *Lexikologie und Lexikographie*, Berlin: AWDDR (Linguistische Studien A 160).

Šipka, Danko. 1992. "Za precizniju klasifikaciju rječničkih etiketa", *Naš jezik*, Beograd, - in print.

Small, Steven, Garrison Cottrell, Michael Tanenhaus (eds.). 1988. *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence.* San Mateo: Morgan Kaufmann Publishers, Inc.

Zgusta, Ladislav. 1988. *Lexicography Today.* An annotated bibliography of the theory of lexicography. (Lexicographica, Series Maior 18), Tübingen: Max Niemeyer Verlag.

Povzetek

MREŽA KVALIFIKATORJEV: PRISTOP K VARIANTNOSTI V SLOVARJU

Kot svoj prispevek k bolj sistematični obravnavi variantnosti v slovarju predlaga pisec izdelavo t.i. mreže kvalifikatorjev in njenega algoritma za obdelavo dvo- in sopomenskosti.