Lexical Decisions in Humans and Computers

Danko Šipka
Slavic Department
Adam Mickiewicz University, Poznan, Poland
e-mail: sipkadan@amu.edu.pl

Lipka Symposium, Munich, Germany, 19-21.6.1998, published in: Falkner & Schimd (ed.) Words, lexemes, concepts - approaches to the lexicon, Gunder Narr Verlag, Tübingen, 1999, p. 391-401.

1. Introduction

In its broadest sense, a decision can be defined as any act of selection among several possible options. In recent decades there has been numerous works both describing and formalizing the entire decision making process. Such works have primarily been accomplished in psychology where conscious decision making is analyzed and in game theory where research is based upon the interrelation and interaction of two or more players. Both fields use such criterion as utility, risk, cost-benefit ratio, etc. More about these approaches can be found in Morrow (1994), Keeney (1996), von Winterveldt and Edwards (1986).

In the use of language, some situations embody conscious decisions involving the person one is communicating with -- the other player as it were. A situation where one must consciously choose an honorific or solidary form of address is a typical example of a game theory type of decision making. In some other situations such decisions are unconscious and fail to take into account the other player, or, to be more precise, the other player is ultimately irrelevant in such decision-making.

Lexical decision can be defined as choosing one from the range of possible lexemes. It is here, much more than in the other fields of the use of language, that the decisions are heuristically rather than algorithmically based. When faced with the high complexity of their environment and lexical decisions (per definition involving high complexity tasks) people often reach to the heuristics and schemata in an effort to reduce the effort required to cognitively process such data (Fiske and Taylor, 1991; Kahneman, Slovic and Tversky 1982).

It can be argued that representing certain lexical choices in the form of heuristically based decision trees might enhance both the learning and computational processing of the languages in question. In this paper I will present some initial results of the project entitled „Decision Heuristics in the Use of Language". The first part of the paper is devoted to the process of choosing L2 lexical equivalent in the Neurotran® MT system. In the second part I discuss the data obtained through text analysis as well as experiments with native speakers of Slavic languages. All research isoriented toward practical applications of the previously mentioned heuristics -- primarily in the fields of computational linguistics and foreign language teaching.

2. Lexical Decisions in Computers

NeuroTran® is a software program developed by Translation Experts Ltd. It is intended to "do things with words" unlike anything that has come before it. It is typically post-fordist and utilizes its knowledge base in various ways depending upon the specific options selected. Its includes a morphological generator and analyzer, a dictionary with lexical list extraction capability, sentence parsing/translation capabilities and quantitative and qualitative analysis. In its sentence translation mode, the software acts as a bilingual transfer system.

The program's principal lexicon task is to choose an appropriate L2 equivalent for each L1 lexeme. Since anisomorphism of lexical systems creates situations like:

$\text{lexeme}_1(\text{L1})=\text{lexeme}_1(\text{L2})/\text{lexeme}_2(\text{L2})/.../\text{lexeme}_n(\text{L2})$

the program in each concrete L1 into L2 translation must choose only one of the possible equivalents.

The choice of equivalents is dependent on actual semantic realization of the lexeme in question. This in turn can partially be inferred from its co*ntext* and *co-text* (see Lipka 1992:24 for more about this differentiation). Consequently, NeuroTran® uses two decision heuristics: a usage labels network to account for the contextual clues and a specific frequency count to account for the co-textual clues.

The usage labels network assumes that the lexemes have the labels stating their valence toward text types. The program first determines the type of text by performing a frequency count of the labeled units and then goes on to choose the lexeme compatible with that particular text type. (Theoretic background of the Usage labels network in: Šipka 1994.) For example, a text about a soccer game will belong to the following label taxonomy:

4	games/plays
44	sport
441	ball games
4411	soccer

and the text about a legal process:

1	sciences
14	social sciences
141	law

If we thus have to solve the following multiple lexical equivalence between Polish and English:

sêdzia,y m;1..44+141/judge n;1..141/referee n;1..44

then in the text concerning a soccer game *referee*, having the label compatible with this respective text type, will be chosen and in the text concerning a legal process, *judge* will be chosen for the same reasons.

Of course, context does not solve all cases of multiple equivalence becausethere is need to address co-textual clues. This is accomplished by giving the status of a specific dictionary entry to each collocation which might reduce multiple lexical equivalence. The program then uses the following algorithm to find the equivalent

If we represent the units of a text as:

1 2 3 4...n

then the process of answering the question „is there an equivalent for" can be presented as in the Table 1:

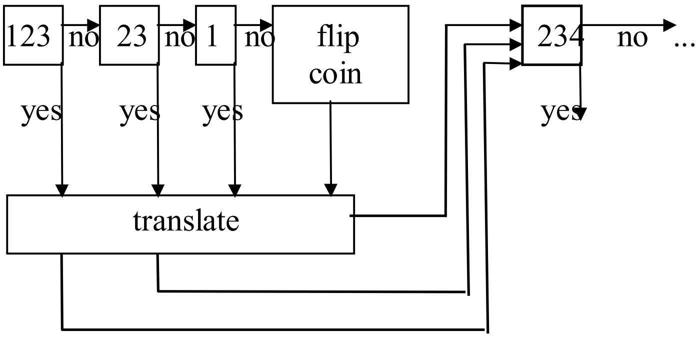


Table 1

Thus if a one-word lexical unit has multiple equivalents, such as in the example:

aktivan,vna,vno--;/actable--aj;/active-aj;/activated--aj;/busy aj;/live--aj;

the procedure finds the multi-word lexical unit first therebyeliminating the problem of multiple equivalence before it ever existed:

aktivni,a,o--; ugljen,a m;/activated charcoal n;

Further means of solving multiple equivalence problem is the use of frequency data.
If, after usage label network and collocation heuristics have eliminated some of the possible equivalents, multiple lexical equivalence possibilities remain as in the case:

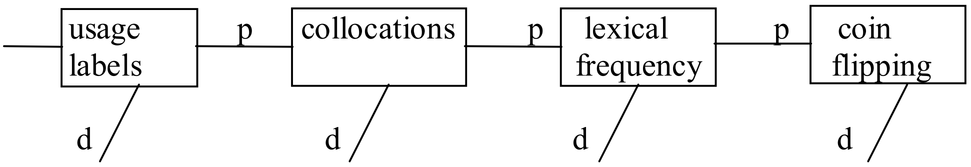
file n; 1...191/datoteka,e f; N1 1...191/datnica,e f; N2 1...191/svež anj,ž nja m; N3 1...191 (N1>N2, N1>N3)

1...191 - lexeme belonging to the computer science field of usage, Nn - frequency data

then the first element ischosen because of its higher frequency.

Finally, if the multiple equivalence still exists, the program can "flip a coin" to decide which equivalent to choose, providing a probability of 0,5 that it will choose the proper equivalent.

We can summarize the process of choosing lexical equivalents in the‘ centipede’ decision tree shown in the Table 2:



p - pass
p - decide

Table 2

Every decision can be changed by the user. The program keeps track of such changes and this data is used as a knowledge base to change the structure of the dictionary thereby providing another means of solving lexical equivalence.

This section of the paper has demonstrated the way in which several lexical decision heuristics can effectivelly be used within a MT system.

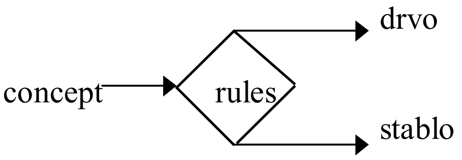
3. Lexical Decisions in Humans

In this section I will discuss two instances of lexical decisions with the speakers of Slavic languages: a.‘ drvo/stablo’ alternation in Serbo-Croatian, and b.‘ ON/IN’ alternation in Polish and Serbo-Croatian.

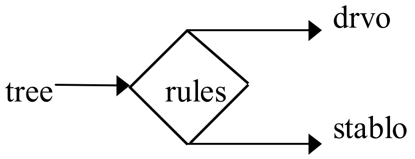
English lexical item *tree*, in its sense:

1. a plant having a permanently woody main stem or trunk, ordinarily growing to a considerable height, and usually developing branches at some distance from the ground. (Word Perfect Corporation, 1994) has two Serbo-Croatian equivalents. That is to say, this conceptual representation can be represented by two Serbo-Croatian lexical units. We thus have the following situations:

a. *Within Serbo-Croatian*



b. *English into Serbo-Croatian*



If we look at the collocations of these two lexemes we can observe that the word‘ drvo’ is used in such contexts were it is construed as an animate object, whereas the word‘ stablo’ can be found in the collocations where animation is not assumed. We thus have: *drvo poznanja* ‘ tree of knowledge’ , *drvo ž ivota*‘ tree of life’ , *nijedno drvo ne raste do neba* ‘ no tree grows up to the sky’ , but: *rodoslovno stablo*, *genealoško stablo* ‘ genealogical tree’ . This animate/inanimate differentiation is further supported by the discussion of the native speakers of Serbo-Croatian on the discussion list ST-L@math.amu.edu.pl, where some of them rejected the idea of using ‘ drvo’ in the Serbo-Croatian translation of ‘ tree diagram’ .

These data provide grounds for the hypothesis that ‘ drvo’ will be preferred in those contexts where tree is construed as something animate. We can further extend this hypothesis and ask if ‘ drvo’ will be preferred in the context where it is construed as something active as well as if it is preferred in the temporal relations (as defined by cognitive linguistics, see Langacker, 1991) as opposed to atemporal relations.

In order to test this hypothesis I have formed a concordance from the Aarhus Serbo-Croatian corpus (available at: ftp aau.dk/pub/slav, ana.yurope.com/pub/books/yu; 4242310 bytes, 728952 tokens, containing contemporary literature in Serbo-Croatian). The following frequency of these two lexical items has been obtained:

drvo	95	64.63%
stablo	52	35.37%
Total	147	100.00%

As the next step, the following table has been created:

A	**B**	**C**	**H**	**context**	**keyword**	**context**
1	1	1	1	izaæutasm o.0 vo	drveæe	nije prijatelj; štiti nas
1	1	1	1	gledam - nema nikog, samo	drveæe	maše.Ostalo je još malo

A - animate (1), inanimate (2)
B - active (1), inactive (2)
C - temporal (1), atemporal relation (2)
H - drvo (1) stablo (2)

Finally, Pearson’ s correlation coefficients (between A,B,C on one hand and H on the other) have been computed. The results are summarized in the Table 3:

	A and H		C and H		B and H
animacy	,2481	temporal	,1995	activity	,1858
hypothesis	(147)	relation	(147)	hypothesis	(147)
	P= ,002	hypothesis	P= ,015		P= ,024

Table 3

The results show that this decision is partially made based on the conceptual representation of the object being animate. However, this is a „soft” type of rule which in all practicality does not have to be addressed in foreign language learning. The point is that there are numerous other predictors of one or another lexical unit and that this one controls only a limited range of the variation. This brings us to a more general conclusion that there are such lexical decisions foreign language students need not be concerned about.

On the contrary, the Slavic ON/IN alternation (defined here as the alternative usage of the prepositions 'na' and 'u'/'w(e)' in the Slavic locative PPs -- depending on the properties of the landmark within the PP and its relationship toward the trajector outside it) belongs to those decisions which cannot be avoided in foreign language teaching.

The ON/IN alternation can be seen in the following examples:

```
            na     + Locative
S-Cr: Ja sam na      krovu.
Pol:  Jestem na      dachu.
      ‘I am on the roof’

           w(e)/u + Locative
S-Cr: Ja sam     u   sobi.
Pol:  Jestem     w   pokoju.
      ‘I am in the room’
```

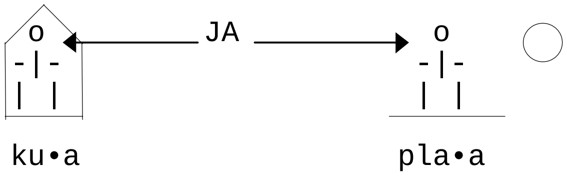
na - ON, u,w(e) - IN, Ja sam, Jestem - I am,, krovu - roof-Loc, dachu - roof-Loc, sobi, pokoju - room-LOC.

In this research, I have taken into account only those locative PPs which pass the "where-test", that is which answer to the question "Where?" This provides only those locative PPs where, in a very broad sense, landmark is a location while at the same time excluding those where landmark is time, circumstance, etc. Finally, I am interested only in the general lexicon excluding any toponymic landmarks.

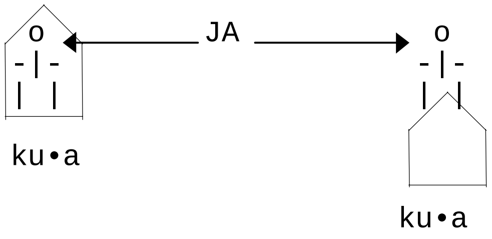
In the trajector-landmark relation expressed by the locative PPs, there are such landmarks which can be construed in that they can be spatially represented (e.g. 'room', 'house', 'tree', etc.) But there are also those which cannot be represented in such a manner (e.g. 'meeting', 'war', 'elections', etc.). This is the first distinction relevant for the IN/ON alternation. Landmark properties can be criteria for this alternation only in those cases where the landmark can be spatially represented. In that case the rule is as follows:

If landmark is mentally represented as an enclosed unit and the trajector is inside this unit, use IN, otherwise use ON. For example:

S-Cr: a. Ja sam u ku•i vs. b. Ja sam na pla•i



S-Cr: c. Ja sam u ku•i vs. d. Ja sam na ku•i



ja = I, u = IN, na = ON, ku•i = house-Loc, pla•i = beach-Loc

This situation, where cognitive linguistics provides relevant explanation only in one segment of the alternation, brings us to the question of what is the best way to organize the decision heuristics In order to answer this question, I have used psycholinguistic and corpus linguistic data. In order to discover if there are differences in processing construable and non-construable landmarks, a simple psycholinguistic experiment has been conducted where reaction time to these two landmarks were measured using two independent groups of subjects with group 0 representing construable landmarks and group 1 non-construable landmarks. The methodology of this experiment is summarized in the Table 4.

Design	Independent matched groups
Subjects	2x30 normal adult subjects, Polish native speakers, undergraduate students, freshmen and juniors
Place	Adam Mickiewicz University, Poznañ
Date	February 27, 1997
Type	Lexical decision task (10 sentences per group)
Independent Variable	0 - construable landmark, 1 - non-construable landmark
Dependent Variable	Reaction time (in hundredths of second)
Software Used	Answer-o-meter (self-made), SPSS 5.0
Examples	000nBy³em __ pagórku.(w/n) 001nBy³em __ konsulatacjach.(w/n)

Table 4

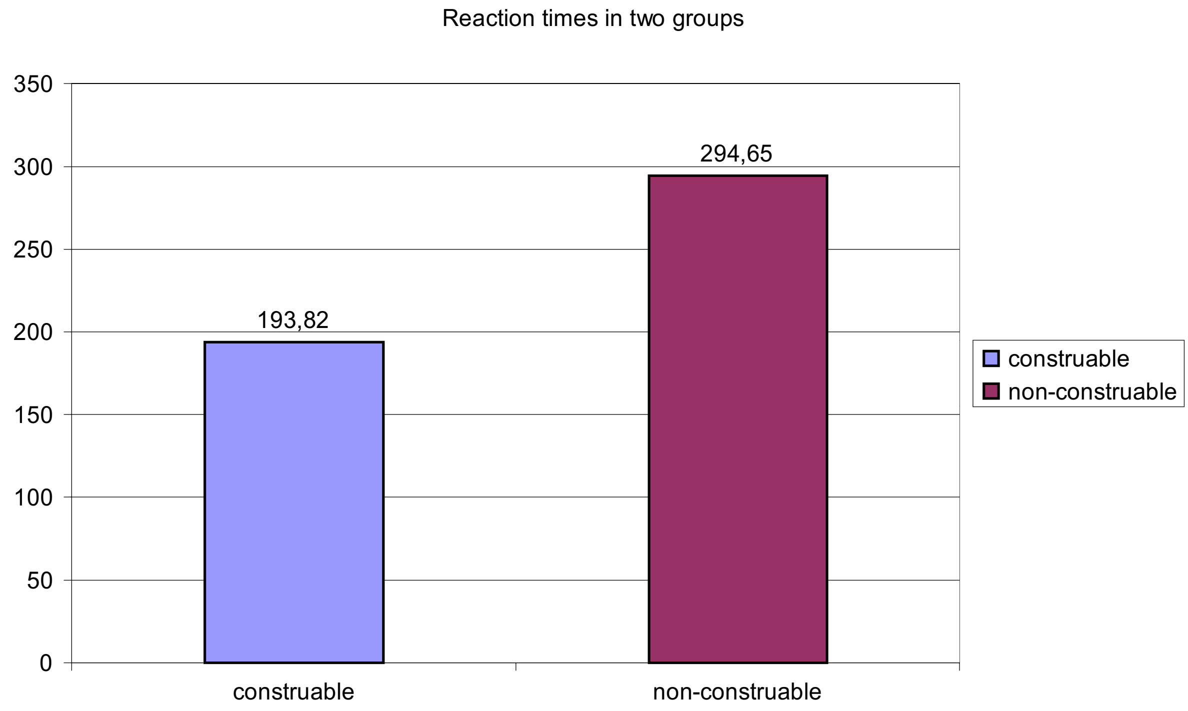
If the distinction between construable and non-construable landmarks is relevant and if some fundamental claims within cognitive linguistics are true, we should expect higher reaction times to the non-construable landmarks.

In order to obtain data about the distribution of this alternation, I have also conducted a quantitative and qualitative contrastive S-Cr - Polish corpus analysis using the texts described in the Table 5:

	Polish	**Serbo-Croat**
Author	Dawid Warszawski	Milan Božić
Content	Articles about the war in the former Yugoslavia	Editorials about the war in the former Yugoslavia
Source	Polish daily *Gazeta Wyborcza*	Serbian radio *B 92*
Form	Electronic document	Electronic document
Period	9/21/93 - 12/7/95	6/3/93 - 2/4/95
Volume	979981 characters	994043 characters
	165566 tokens	193890 tokens
	42914 types	52502 types
Language	Polish, standard, journalist	S-Cr, standard (Serbian variant),journalist

Table 5

The psycholinguistic experiment has revealed that there is a significant difference in processing construable and non-construable landmarks. Descriptive statistical data show that the subjects needed more time to process non-construable landmarks and inferential statistics show that this difference is meaningful. This can be seen in Table 6.



F=32,9412 p<0,01

Table 6

The quantitative contrastive study of two corpora provides the data shown in the Table 7:

	Polish			**Serbo-Croat**		
	#	%		#	%	
total	8500	100%		7820	100%	
na	2827	33%		3017	39%	
w(e)/u	5673	67%		4803	61%	
			% (total)			*%(total)*
con.+non-con.	1739	100%	20%	2040	100%	26%
construable	834	48%		599	29%	
non-construable	905	52%		1441	71%	
difference	100	6%		105	5%	
construable	834	100%		599	100%	
na	377	45%		312	52%	
w(e)	457	55%		287	48%	
non-construable	905	100%	*difference*	1441	100%	*difference*
na	168	19%	1 (1%)	512	36%	104 (20%)
w(e)/u	737	81%	99 (13%)	929	64%	1 (0%)

Table 7

The spatial IN/ON alternation in Locative PPs (as defined here) covers only a small portion (20-26%) of the distribution of these prepositions. But if we include this alternation in the accusative as well as toponyms, we can expect to see that it extends to more than one half of the whole. It can be observed that S-Cr has a higher non-construable to construable ratio than Polish. This is possibly caused by the fact that theauthor of the S-Cr material was primarily interested in political speculation whereas the Polish author was more concerned with pragmatic, concrete events.

However, it is clear that this distinction is such that both options are possibilities. It is not a situation where one is a "default" and the other an "exception". Essentially, we have an 'either A or B' relation, not an 'unless B, A' relation. It is interesting to note that the only discernible differences between Polish and S-Cr were found with non-construable landmarks -- only a small percentage of the total distribution. Furthermore, in most cases, the Polish 'w(e)' and the S-Cr 'na', help explain the difference in the distribution of these prepositions in two corpora. These different forms belonged to 17 different lexemes.

The qualitative analysis revealed that there is no single decision-making criterion as is the case with construable landmarks. The possible decision-making tree in this segment would contain a great number of nodes and exceptions. The alternative is to have only one node and a lengthy list of exceptions. A series of experiments is required to determine which of the solutions is more practical. The qualitative analysis of S-Cr - Polish differences demonstrate that for some of these items one can find common hyperonyms, for example, electronic media (S-Cr: ‘ na televiziji’ - Pol: w ‘ telewizji’ , S-Cr ‘ na radiju’ , Pol: ‘ w radiu’), and politics (S-Cr: ‘ na referendumu’ - Pol: ‘ w referendum’ , S-Cr: ‘ na izborima’ , Pol: ‘ w wyborach’).

If we choose the one-node-with-list model for the non-construable landmarks, the decision-making flowchart (where we one must choose between IN and AT) can be presented as follows in the Table 8 (the areas of S-Cr - Polish differences are shaded):

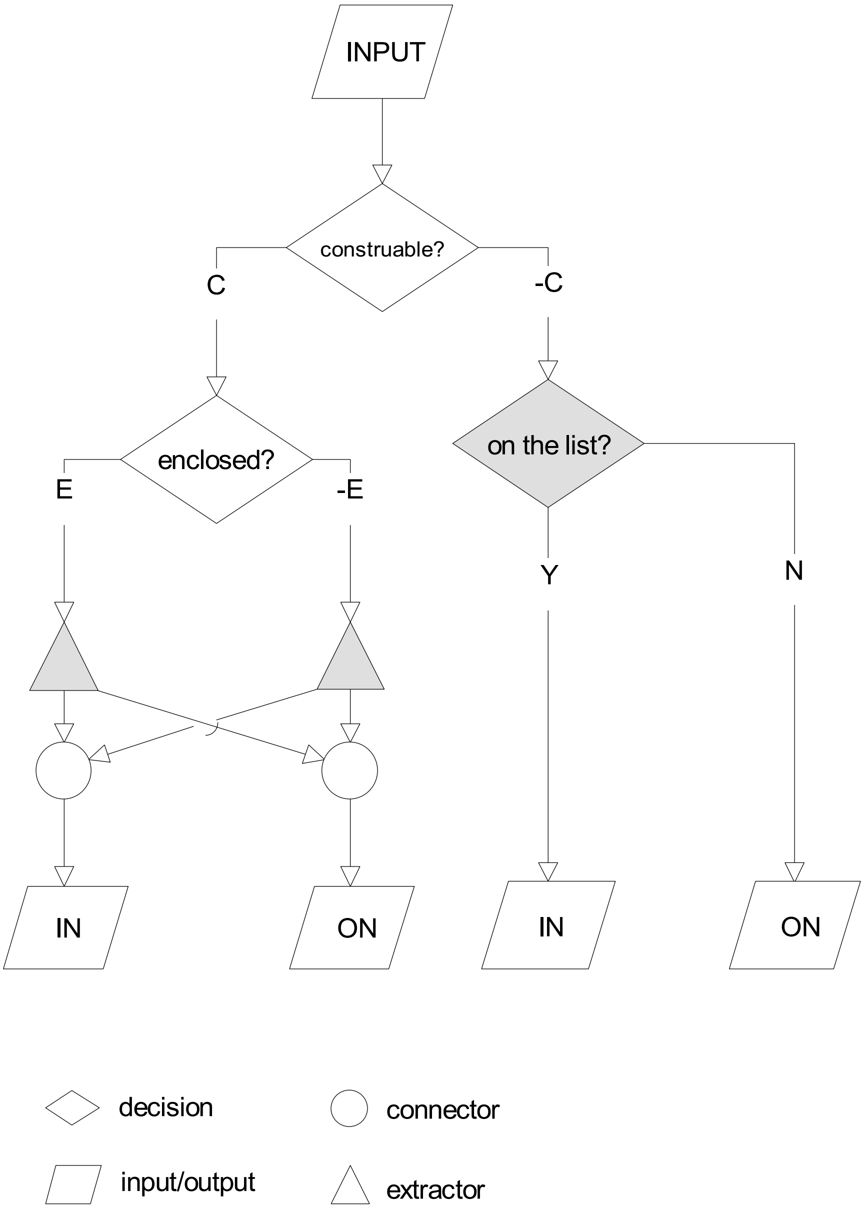


Table 8

In those cases where the landmark is construable, the chart can account for practically all cases. The extractors are reserved for rare cases, primarily place names such as the S-Cr 'Cetinje' or ' Pale' , which take the ' na' form rather than the ' u' form.

The above tree can be used as an effective heuristics for this decision in the foreign language learning. The main rule solves the most of the variations, and in the other cases students gradually acquire the exceptions from the list in the non-construable segment of the decision tree.

4. Directions for Further Research

Although substantially different in many ways, lexical decisions in humans and computers have something in common -- in some situations both computers and humans can profit from heuristics which are not perfect but which are capable of providing a correct solution in the majority of situations. Further work in both areas is planned to determine the best way of organizing these heuristics.

References

Fiske S.T., Taylor S.E. 1991 *Social cognition*, New York: McGraw-Hill

Kahneman D., Slovic P., Tversky A. 1982 *Judgment under uncertainty: Heuristics and biases*, New York: Cambridge University Press.

Keeney, R. L. 1996. *Value-focused thinking. A Path to Creative Decisiomaking*. Harvard: Harvard University Press

Langacker, R.W. 1990. *Concept, Image, Symbol: The Cognitive Basis of Grammar*, Berlin - New York: Mouton de Gruyter

Lipka, L 1992 *An Outline of English Lexicology*, Tübingen: Niemeyer

Morrow James D. *Game Theory for Political Scientists*, Princeton, 1994

Šipka, D. 1994 „Usage Labels Network: An Approach to Lexical Variation", in: *Linguistica*, XXXIV,2:31-42

Word Perfect Corporation 1994 *Random House Webster' s Electronic Dictionary and Thesaurus*, College Edition

von Winterveldt, D. and W. Edwards. 1986 *Decision Analysis and Behavioral Research*. Cambridge: Cambridge University Press