

SemCAPTCHA – the user-friendly alternative for OCR-based CAPTCHA systems[†]

Paweł Łupkowski and Mariusz Urbański

Chair of Logic and Cognitive Science
Institute of Psychology, Adam Mickiewicz University, Poznań, Poland
{Pawel.Lupkowski, Mariusz.Urbanski}@amu.edu.pl

ABSTRACT

In this paper we present a new CAPTCHA system (*Completely Automated Turing Test To Tell Computers and Humans Apart*). The proposal of SemCAPTCHA is motivated by an increasing number of broken OCR-based CAPTCHA systems and it is based not only on text recognition but also on text understanding. We describe both, the user's and the system's perspective of SemCAPTCHA and compare it with some currently popular CAPTCHAs. We also briefly describe an experiment carried out to test our CAPTCHA on human users.

1. Introduction

In many domains, there is an increasing demand for a simple and efficient way to differentiate real human users from malicious programs (bots). A few examples of such domains are: services offering free e-mail accounts, community portals, online polls etc.

One of the most popular ways to tell human users and bot users apart are the so called CAPTCHA systems (this acronym stands for *Completely Automated Turing Test To Tell Computers and Humans Apart* – cf. [1], [2]).

The design of an effective CAPTCHA system is a difficult task, since two distant needs must be satisfied: it has to be really hard for a machine and at the same time it should not decrease the usability of the system. Usability aspects are very important as CAPTCHAs cannot engage too much of the user's attention and cannot consume too much of her time. Registering a free e-mail account is a good example here. There are many alternative providers of such accounts on the market, so if you want a potential user to solve a CAPTCHA on your site, it has to be as unproblematic for her as possible (and you want her to solve it in order to prove that she is a human, not a bot who will send tons of spam from your servers). If a potential user gets irritated, she will go away and pick another provider. To make things more difficult, there's also a third factor: a CAPTCHA has to be open, that is, the algorithms used by a system must be public. The idea is, that CAPTCHA effectiveness should be based on difficulty of an underlying AI problem and not on a secret cryptographic mechanism or other copyrighted mystery. Finally, test instances of a CAPTCHA should be generated automatically.

[†]This research was partially supported by AMU Faculty of Social Sciences grant No. WSO/133/2006.

Internet users encounter CAPTCHAs very often. Most of them are visual CAPTCHAs, where the task consists in recognition of a word or string of symbols (letters, numbers) from a distorted picture. To solve such a CAPTCHA the user has to write down words or symbols from the picture. Such systems work e.g. on Yahoo (<http://www.yahoo.com>), Gmail (<http://www.gmail.com>), Wirtualna Polska (<http://www.wp.pl>), Gazeta.pl (<http://www.gazeta.pl>) and many other sites. Exemplary CAPTCHAs are presented in Table 5.

Currently it is an important issue, that the AI problem underlying such CAPTCHAs is challenged by constantly developing Optical Character Recognition (OCR) systems with an increasing success rate. Mori and Malik [9] describe an attack on a visual CAPTCHA EZ-Gimpy used by Yahoo!, which enjoyed a success rate of 92%. In the more difficult case of Gimpy they passed the test 33% of the time. As the authors claim: “with our 33% accuracy, this CAPTCHA would be ineffective in applications such as screening out “bots” since a computer could flood the application with thousands of requests.” [9, p. 7]. After all, year 2008 seems to be a really bad year for visual CAPTCHAs: Yahoo! CAPTCHA was hacked again (<http://osnews.pl>, 21.01.2008), as well as the Gmail one (<http://osnews.pl>, 27.02.2008), and MS Windows Live Hotmail (<http://arstechnica.com>, 15.04.2008). Many visual CAPTCHAs are broken ‘out of the box’ by the PWNtcha system (see <http://libcaca.zoy.org/wiki/PWNtcha> – examples of 12 broken CAPTCHAs where the success rate is from 49% to 100%).

As a consequence, there is a great need for more secure alternative CAPTCHAs, which are based not only on the OCR problem. There are some proposals, like the question-based CAPTCHA [7], ARTiFACIAL [13], PIX [1], sound oriented CAPTCHAs [3] etc. In our opinion, the current situation offers a great motivation to look for an inspiration for CAPTCHA systems not only in simple sensory processing but also in higher levels of human data processing.

2. The SemCAPTCHA system

Our proposal is to base a CAPTCHA system on a combination of an OCR problem and some linguistic task, and to apply the effect of positive semantic priming to strengthen the humans’ odds against the computers’. Let us first remind what semantic priming is:

Priming is an improvement in performance in a perceptual or cognitive task, relative to an appropriate baseline, produced by context or prior experience. Semantic priming refers to the improvement in speed or accuracy to respond to a stimulus, such as a word or a picture, when it is preceded by a semantically related stimulus (e.g., *cat-dog* relative to when it is preceded by a semantically unrelated stimulus (e.g., *table-dog*). The stimulus to which responses are made (e.g., *dog*) is the target and the preceding stimulus (e.g., *cat* or *table*) is the prime. [8, p. 3–4]

All that is needed to break a simple visual CAPTCHA is a good OCR program. Breaking our system – SemCAPTCHA, where “Sem” stands for “semantic” – is not that straightforward for a machine, although – as we intend to achieve – it should still remain quite simple for a human user.



Figure 1. Sample instance of SemCAPTCHA test.

2.1. SemCAPTCHA – the user’s perspective

A SemCAPTCHA test instance consists of a distorted picture, on which three words are presented. One word differs from the other two in its meaning. The task is to recognize this word and point it by a mouse click. It has to be stressed that the words do not differ substantially as for their graphical properties (like, e.g. length): the difference is of semantic character. We have designed SemCAPTCHA to work on names of animals. In our experiment one animal was different from the rest in that it was, e.g., a mammal among reptiles. Alternative designs are possible as well.

An example of such test instance is given in Figure 1: a user is presented with the words “kaczka” (a duck), “kukułka” (a cuckoo), “krowa” (a cow; SemCAPTCHA is designed in Polish). The proper answer is “krowa” and the semantic difference is based on taxonomy: ducks and cuckoos are birds while cows are mammals.

To solve this task, the user has to first recognize the words from a distorted picture, then identify their meaning and finally find an underlying pattern and the word which does not fit it. The choice of words makes it easy even for not very fluent language users.

In order to make SemCAPTCHA even easier for humans we decided to employ the positive semantic priming effect. Each test instance is preceded by a prime (the sequence is prime – mask – test instance, where the mask is for example a string of X). The prime is a word semantically connected with the task solution; in the case of the example presented in Figure 1 it might be the word “mleko” (milk). This setting should enable a human user to recognise the answer much faster than in the case when there’s no prime. Consequently, a human user should solve SemCAPTCHA test instances easier and faster (cf. next section, [5] and [6]).

2.2. SemCAPTCHA – the system’s perspective

SemCAPTCHA has not been implemented yet, but the procedures needed for the system are under development. Because of this we may present only a sketch of future solutions for SemCAPTCHA system.

SemCAPTCHA would work on a word base consisting of about 500 animals’ names. Names would be grouped in categories, e.g. mammals, birds, reptiles. Each word would have its own semantic field (stored as the assertional semantic network). The semantic field would contain words semantically connected with a given animal name (this solution is based upon methods presented in [14] and [15]). Each connection of words would be marked by a label containing information about the relation between respective objects (e.g. cow – gives – milk) and the relation’s strength. The strength would be expressed by a numerical value 1–100 (calculated on the basis of how often those words

appear together in IPI PAN – corpus of Polish developed by the Polish Academy of Sciences – and Google). We think that such an architecture would enable efficient and automatic generation of test instances.

This procedure might work as follows. To generate a test instance the system randomly chooses two categories from the word base. First, it picks (also randomly) one word (w_1) from one category and two words from another one (w_2, w_3). Second, the system picks a prime for w_1 , using the semantic network stored for w_1 . The system randomly chooses possible relation strength with w_1 of certain range (e.g. 50–70) and a word that obeys this restriction. Third, a distorted picture is generated using w_1, w_2, w_3 and it is preceded by a prime and a mask.

After a test is generated and displayed, SemCAPTCHA starts to measure the time. The solution time (the interval between exposition of a picture and a mouse click) is compared with a standard solution time for SemCAPTCHA. Our experiment shows, that for humans the solution time varies from 1,2 to 5,5 seconds (cf. next section, [5] and [6]; more thorough research could help to verify these limits). This is one of the most characteristic properties of SemCAPTCHA: it not only generates and evaluates test instances but it also constantly records the solution time, and its verdict depends not only on the correctness of a solution but also on the time needed for it. In this aspect SemCAPTCHA differs substantially from the widely used OCR-based CAPTCHA systems.¹

3. The SemCAPTCHA experiment

To verify the idea of using linguistic competence and positive semantic priming in the SemCAPTCHA system, we have carried out an experiment (details on the instruments used and methods of statistical analysis can be found in [5], [6] and are available from the authors).

Our research questions for these issues were:

- 1) Is the effect of positive semantic priming statistically significant for the solution time of SemCAPTCHA test instances?
- 2) Is the effect of positive semantic priming statistically significant for solution accuracy of SemCAPTCHA test instances?

The experiment consisted of one training task and 10 test instances. A single instance consisted of a picture with 3 Polish words (names of animals). One word was different from the other two in that it was a name of an animal of a different class. For each picture we used one of the standard CAPTCHA's methods of distortion. We prepared two sets of tasks, *A* and *B*, consisting of the same test instances. In the experimental set *A*, each test instance was preceded by a prime, semantically connected with the word which was the correct solution of a task. The prime was followed by a mask. In the control set *B* there was no prime. Detailed characteristics of test instances are given in Table 1.

¹We would like to give our thanks to Maciej Piasecki, who pointed out that this solution may lead to some problems, since CAPTCHA's algorithm should be public. Thus there will be possible for a malicious bot to simulate estimated solution time for SemCAPTCHA tasks. We should however remember that SemCAPTCHA's hardness assumption relays not only on the time comparison mechanism, but also on other issues, like OCR problem and linguistic competence.

Table 1. Tasks characteristics

Task	Prime (ms)	Mask (ms)	Text dist.	Bg. dist.
T1	70	50	G-blur	HSV
T2	60	50	G-blur	RGB
T3	80	50	G-blur	fog
T4	90	50	dispersion	HSV
T5	100	60	dispersion	RGB
T6	60	30	dispersion	fog
T7	70	50	Whirl&Pinch	HSV
T8	70	50	Whirl&Pinch	fog
T9	70	50	Whirl&Pinch	RGB
T10	70	50	newspaper printout	HSV

The sample consisted of 64 students at the Adam Mickiewicz University (19 males, 43 females, 2 no data; average age 21),² who volunteered to participate in the experiment. The participants were randomly divided into groups *A* (experimental group) and *B* (control group).

The subjects were asked to choose in each picture from three names of animals the name of an animal which differs from the other two and point it by a mouse click. Solution time was measured as an interval between the exposition of a picture and the click. Time and correctness of the solution were recorded down automatically by the server. After the completion of all ten test instances the subjects were asked to fill a short questionnaire concerning subjective difficulty of each task (a complete set of pictures was presented on the monitor at this stage) and their willingness to solve such tasks while surfing the Internet.

The results enabled us to formulate a positive answer to our first question and a negative answer to the second one (cf. Table 2). First and foremost, we observed the effect of positive semantic priming in solving test instances of SemCAPTCHA: there was a statistically significant difference in the time of solving test instances between the experimental group (*A*) and the control group (*B*). The participants from group *A* solved test instances faster than the participants from group *B* and thus it is possible to differentiate between the experimental and the control group on the basis of the average time of solving test instances. This effect was present in the case of eighth out of ten test instances (T3–T10). The lack of positive semantic priming effect in the case of the first and second instance can be explained by the need for some practice in solving such tasks.

On the other hand, the improvement in the time of solving test instances does not affect, in a statistically significant way, the accuracy of the solutions. The participants from the experimental group solved test instances faster, however, not more accurately than participants from the control group.

²According to research performed by Cracow University of Economics 32% of Polish Internet users are between 21–25 years old (cf. <http://www.badanie.ae.krakow.pl>). Those results were confirmed by research performed by I-Metria (cf. <http://www.cxo.pl/news/44192.html>).

Table 2. Average time, accuracy and subjective difficulty of task solutions

Task	Group	N	Average time (sec.)	Accuracy	Difficulty (Average)
T1	A	31	5,5408	17	6,35
	B	33	5,9048	16	6,55
T2	A	31	2,3467	30	2,61
	B	33	2,7859	32	3,56
T3	A	31	1,8594	27	3,26
	B	33	2,7749	31	3,48
T4	A	31	2,7456	21	5,61
	B	33	4,7085	25	6,50
T5	A	31	1,2047	31	3,00
	B	33	3,3308	31	3,63
T6	A	31	1,8863	30	3,03
	B	33	2,8534	32	3,47
T7	A	31	2,5314	21	4,50
	B	33	3,5239	22	5,28
T8	A	31	1,7810	28	3,67
	B	33	3,2051	31	5,25
T9	A	31	1,4193	30	2,67
	B	33	2,6340	32	2,97
T10	A	31	1,5180	23	3,07
	B	33	2,6648	27	3,47

4. SemCAPTCHA and other proposals

As we have already pointed out before, user friendliness is one of the crucial issues for an effective CAPTCHA system: for humans they should be as easy as possible. Thus, it is interesting to compare our system with other CAPTCHAs on the basis of a declared subjective difficulty of test instances and a declared willingness to use them in practice. We have chosen CAPTCHAs for which such data was available for comparison.

We have already mentioned that in our experiment we asked participants to declare subjective difficulty of test instances (on the scale 1–10, where 1 means the simplest). For each test instance the subjective difficulty declared by the participants from the experimental group was slightly lower than the one declared by participants from the control group (however, only in one instance this difference was statistically significant). We observed high correlation between the average declared difficulty and the average solution time (for group A $r^2 = 0.71$). As a consequence, the time of solution seems to be a good estimator of task complexity. This observation gives some basis for comparing SemCAPTCHA with other CAPTCHA systems on the objective basis of their solution times.

One of the alternatives for OCR-based CAPTCHA is ARTiFACIAL (cf. [13]). It is based on the ability to recognize faces. The motivation for this system is quite similar to ours – make use of higher levels of human data processing. The ARTiFACIAL test consists of one picture containing background (with randomly chosen facial features)

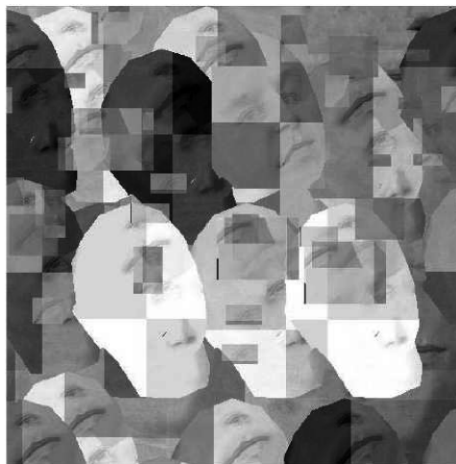


Figure 2. Example of ARTiFACIAL test [13].

Table 3. Average time (in sec.) of ARTiFACIAL test solution

task	1	2	3	4	5	6	7	8	9	10
time	22	15	16	13	12	11	12	12	11	12

and a face (exemplary test instance is presented in Figure 2). The task is to find and point six points in such a picture (left and right corner of: left eye, right eye and mouth). As could be expected, ARTiFACIAL is really hard for machines, but is it simple enough for human users? ARTiFACIAL authors carried out an experiment on this issue. It consisted of 10 ARTiFACIAL test instances. The sample consisted of 34 subjects (accountants, administrative staff, architects, executives, receptionists, researchers, software developers, support engineers, and patent attorneys). Average solution times are presented in Table 3 (cf. [13, p. 500]).

The mechanics of ARTiFACIAL and SemCAPTCHA are quite similar and it can be claimed that ARTiFACIAL's underlying problem is not more difficult than the SemCAPTCHA's one. Thus, if we use the solution time as an estimator of task complexity for human users we may say that ARTiFACIAL is a really complex CAPTCHA system. The average solution time for all tasks is 14 seconds. SemCAPTCHA seems to be much easier, since the average solution time is 2.3 seconds (cf. Figure 3).

On the basis of the declared willingness to use them in practice we can compare SemCAPTCHA to a simple visual CAPTCHA system – BaffleText. [4, p. 7] presents the results of a short questionnaire which was ment to investigate BaffleText users' feelings about this system. It has been filled by 18 out of 33 subjects (Palo Alto Research Center employees):

- 1) 16,7% reported they would be willing to solve a BaffleText each time they sent email;
- 2) 38,9% reported they would be willing to solve a BaffleText, if it reduced spam tenfold;

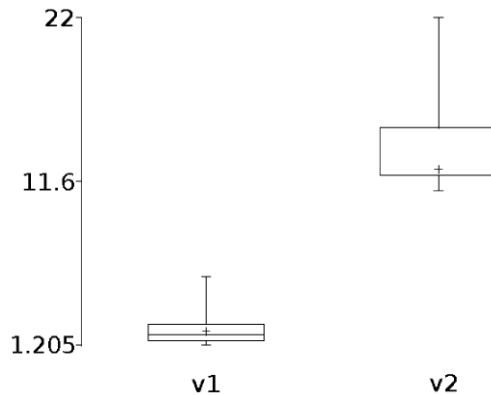


Figure 3. Average time of solution for SemCAPTCHA (v1) and ARTiFACIAL (v2).

- 3) 94,4% reported they would be willing to solve a BaffleText, if it meant those sites had more trustworthy recommendations data;
- 4) 100% reported they would be willing to solve a BaffleText each time they registered for an e-mail account.

In our experiment we asked subjects to answer the same questions (61 out of 64 did this):

- 1) 15,6% reported they would be willing to solve a SemCAPTCHA each time they sent email;
- 2) 43,8% reported they would be willing to solve a SemCAPTCHA, if it reduced spam tenfold;
- 3) 65,6% reported they would be willing to solve a SemCAPTCHA, if it meant those sites had more trustworthy recommendations data;
- 4) 34,4% reported they would be willing to solve a SemCAPTCHA every time they registered for an e-mail account.

We think that these results are very promising for SemCAPTCHA. One possible explanation of such results for the third and fourth question is that our subjects were students. They might be might not be so keen on web security issues as PARC employees.

We have also performed some OCR tests, to see how hard SemCAPTCHA tests for OCR programs are. SemCAPTCHA uses slightly distorted pictures, so we intended to compare them with OCR-based CAPTCHAs currently used by popular portals. We tested our experimental test instances against three OCR programs: GOCR, Asprise OCR and ABBYY Fine Reader 9.0 PE. The results (the percentage of correctly recognized words and symbols) are presented in Table 4.

For comparison we also performed OCR tests (against the same three programs) for other popular visual CAPTCHAs: the ones used by Yahoo!, wp.pl and gazeta.pl (10 instances for each). These CAPTCHAs do not use regular words, but only strings of symbols (letters and numbers). Exemplary tasks are presented in Table 5.

For the CAPTCHA used by Yahoo! (considered as difficult) GOCR recognised 2.82% of the signs; Asprise OCR 1.41% and ABBYY FR 19.72%. As for wp.pl results were following: GOCR 52.94%, Asprise OCR 16.67%, ABBYY FR 5%. And for gazeta.pl: GOCR 45%, Asprise OCR 0%, ABBYY FR 47.06%. All results are presented in Figure 4.

Table 4. OCR tests for semCAPTCHA

GOCR		Asprise OCR		ABBYY FR	
words	letters	words	letters	words	letters
0%	4.11%	0%	6.16%	13.33%	13.01%

Table 5. Exemplary tasks of captchas used by Yahoo!, wp.pl and gazeta.pl

Yahoo!	wp.pl	gazeta.pl

All tested CAPTCHAs are based on the OCR problem. The SemCAPTCHA results are comparable with the others (and it should be stressed that recognising the words in the SemCAPTCHA task is only the first step towards the solution; cf. section 2). Thus we may conclude, that the OCR-hardness of SemCAPTCHA is set high enough (i.e. it is at least as hard for machines as the CAPTCHAs we have tested and which are currently used by leading internet portals).

5. Conclusions

SemCAPTCHA, based on a combination of an OCR problem, some linguistic task and positive semantic priming, seems to be a promising system for telling humans and computers apart. On one hand, the engagement of higher level human data processing makes it harder for machines than the currently used visual CAPTCHAs. On the other

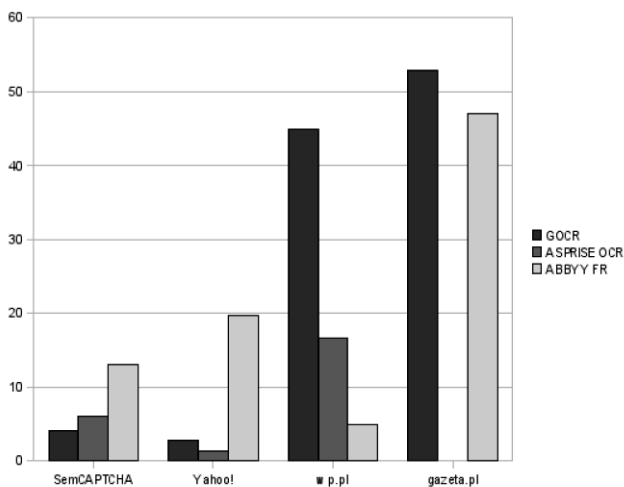


Figure 4. OCR tests results (in % of recognised symbols).

hand, it is not as complex for human users as other alternatives to the current systems. SemCAPTCHA has a simple and open algorithm, it is easy for humans and can be designed for any language.

BIBLIOGRAPHY

- [1] Ahn L., Blum M., Hopper N. J., Langford J. CAPTCHA: Using Hard AI Problems For Security. Retrieved October 11, 2007 from <http://www.captcha.net>.
- [2] Ahn L., Blum M., Langford J. Telling Humans and Computers Apart Automatically. How Lazy Cryptographers do AI. Retrieved October 11, 2007 from <http://www.captcha.net>.
- [3] Chan N. Sound oriented CAPTCHA. Retrieved October 11, 2007 from <http://www.captcha.net>.
- [4] Chew M, Baird H. S. (2003). BaffleText: a Human Interactive Proof. Proceedings of the SPIE/IS&T Document Recognition and Retrieval Conf. X. Santa Clara, CA.
- [5] Łupkowski P., Urbański M. (2006). Positive semantic priming optimization tool for automated user authorization systems. Research report, Institute of Psychology, Adam Mickiewicz University (in Polish).
- [6] Łupkowski P., Urbański M. (2008). SemCAPTCHA. Telling Computers and Humans Apart by Means of Linguistic Competence and Positive Semantic Priming, In L. Rutkowski, R. Tadeusiewicz, L. A. Zadeh, J. Zurada (Eds.), Computational Intelligence: Methods and Applications (pp. 525–531). Academic Publishing House EXIT.
- [7] Shirali-Shahreza M., Shirali-Shahreza J. (2007). Question-Based CAPTCHA, Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007) – Volume 04 (pp. 54–58). IEEE Computer Society, Washington DC.
- [8] McNamara T. P. (2005). Semantic Priming: Perspectives from Memory and Word Recognition. Psychology Press. Taylor & Francis Group. New York.
- [9] Mori G., Malik, J. (2003). Recognizing objects in adversarial clutter: breaking a visual CAPTCHA. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2003. Retrieved October 11, 2007 from <http://www.cs.sfu.ca/~mori/research>.
- [10] Naor M. (1996). Verification of a human in the loop or Identification via The Turing Test. <http://www.wisdom.weizmann.ac.il/~naor/PAPERS/human.ps>.
- [11] Neely J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner, & G. W. Humphreys (Eds.), Basic processes in reading (pp. 264–336). Hillsdale, NJ: Lawrence Erlbaum Associates.
- [12] Plaut D. C. (1995). Semantic and Associative Priming in a Distributed Attractor Network. In Proceedings of the 17th Annual Conference of the Cognitive Science Society (pp. 37–42). Hillsdale, NJ: Lawrence Erlbaum Associates.
- [13] Rui Y., Liu Z. (2004). ARTiFACIAL: Automated Reverse Turing test using FACIAL features. *Multimedia System* 9: 493–502.
- [14] Szymański J., Sarnatowicz T., (2005). Concept description vectors and the 20 question game, In: Intelligent Information Processing and Web Mining, Eds. M.A. Kłopotek, S.T. Wierzchon, K. Trojanowski, Advances in Soft Computing (pp. 41–50). Springer Verlag.
- [15] Szymański J., (2006). Pamięć semantyczna i awatar grający w 20 pytań (Semantic memory and Avatar for 20 questions game), 2nd Polish and International PD Forum-Conference on Computer Science, Smardzewice-Łódź, 16–19 October 2006 (In Polish).