Acta Cybernetica **18** (2007) 329–340.

How to Represent Meanings in an Ontology *

Miklós Szőts[†], Tamás Grőbler[†] and András Simonyi [†]

Abstract

We work on a method for giving a formal semantic representation of natural language texts. The semantic representation is generated in an ontology, on the basis of morphological and syntactic information. The task of the semantic analysis is to create instances in the ontology that contains the world model, i.e. to create those individuals and relations that correspond to the situation described by the text. The knowledge base of the semantic analyser is stored in an OWL ontology. This paper gives an overview of the system, and we discuss those questions of ontology design that require special attention in the context of meaning representation. We also present a software prototype that is based on the method and generates electronic medical records from free-form medical texts.

Keywords: ontology, semantics, natural language processing, electronic medical record

1 Introduction

The research presented here aims at giving a formal representation of the semantic content of natural language texts. A method for formal meaning representation can be put into use in many areas—we have already implemented a software prototype that generates formally structured medical records from free-form texts, but our most important long-term goal is to develop a semantic search engine (text mining tool).

Currently we handle only the descriptive function of language, i.e. we represent only the meaning of declarative sentences. Our approach is based on the idea that the meaning of a text (or text fragment) is the representation of the described situation in a world model.

*This work was supported by research grants GVOP-3.1.1-2004-05-0363/3.0 and NKFP-2/042/04 (MEO).

[†]Applied Logic Laboratory, E-mail: {szots,grobler,simonyi}@all.hu

The method uses a knowledge base consisting of the following modules:

- knowledge about the world, that is, the world model;
- linguistic knowledge,
- a mapping representing the connections between these components.

Both the world model and linguistic knowledge are represented in ontologies. Using these modules we are able to store and use the knowledge that is required for interpreting natural language texts: an algorithm can be formulated that tracks the referential connections between expressions of a given text and elements of the world model on the basis of the above mentioned mapping, and 'populates' the world model with the situations, objects, relations etc. that are described. These additions to the world model ontology are, in effect, formal representations, that can be collectively described as the meaning of the text in question.

The paper is organised as follows. Section 2 surveys issues connected with the world model, while Section 3 gives an overview of the linguistic knowledge base, its relation to the world model, and, most importantly, the method of using the three modules for generating semantic representations. The remaining four sections contain a description of the 'proof of concept' software prototype we have implemented, a short discussion of related work, and a summary of our plans for further development.

2 The world model

In order to be able to represent the meaning of texts about a given area of knowledge, the ontology has to be capable of representing those situations that typically occur in the texts to be analysed. Consequently, it has to contain both domain-specific concepts, and concepts corresponding to everyday words that connect the domain-specific expressions with each other. This requires, first of all, a satisfactory top-level ontology, which determines what is expressible in the system. In the following we discuss some of the most important issues that has to be faced when designing a top-level ontology. We raise these problems on a general level, but it is to be emphasised that the categories of a special domain ontology might be radically different from those of general purpose ontologies. For instance, it is totally unnecessary to include the common genus HUMAN of the concepts PATIENT and MEDICAL_STAFF_MEMBER in a medical ontology.

2.1 Eventualities

Eventualities (also known as occurrences or perdurants in the literature: they include both events and states) are of crucial importance for natural language processing, since they are those elements of reality (or our representation thereof) that are usually referred to by verbs. We have introduced the relation PARTICIPATES_IN between the concepts ENDURANT and EVENTUALITY (the former concept applies

to all physical or abstract objects that persist through time), and thematic roles are considered to be subrelations of it, determining those concepts whose instances can be e.g. the actor, the object etc. of an event—see [15].

2.2 Properties

How should we represent that 'the sky is blue' i.e. that 'the colour of the sky is blue'? Or, for that matter, how should we represent the sentence that 'on the 8th of November, 2006, the patient's blood pressure was 220/178 mmHg, measured on her left arm, when she was sitting'? Although the first example seems to be simple—even if there is a hidden time dependence—the second illustrates an obvious problem: we have to represent the fact that an instance of the concept PATIENT has a property (BLOOD_PRESSURE) with a given value ('the patient's blood pressure') which depends on various parameters (position, place of measurement, time). Obviously, different properties will depend on different parameters. We chose to solve this problem by reifying the relation 'has property,' and to connect the reified, individualised properties (also known as tropes in the philosophy literature) with the relations BEARER, HAS_VALUE, and relations corresponding to the parameters (IN_POSITION, HAS_PLACE_OF_MEASUREMENT, HAS_TIME in the example) to the relevant objects and values.

2.3 Time

Almost every domain's representation requires a representation of time, and we opted for a relatively simple treatment: a distinction is made between time intervals and time points, and the class of time points is mapped onto a linear scale. We introduced the HAS_START and HAS_END relations with the concepts TIME_INTERVAL as their domain and TIME_POINT as their range. Unfortunately, this simple picture is spoilt by the problem of granularity: time expressions, like *day*, or *month* can refer either to time intervals or time points, depending on the context. This problem can be resolved in a number of ways—we chose to take the referents of these expressions exclusively as time points that can be the beginning or end of certain time intervals. It is the task of the semantic analyser to find those time expressions that in fact refer to time points serving as endpoints of intervals.

2.4 Location

The representation of locations is essentially different from that of time points and intervals, since there is no unified, common sense coordinate system for them. Consequently, we have to use other concepts to determine locations. In medical contexts we encounter two, totally different ways of referring to places: certain medical concepts are connected to body parts (e.g. a liver tumour), while in other cases medical units (hospitals etc.) are the locations that are referred to. We represent these two ways of locating an object by two different relations.

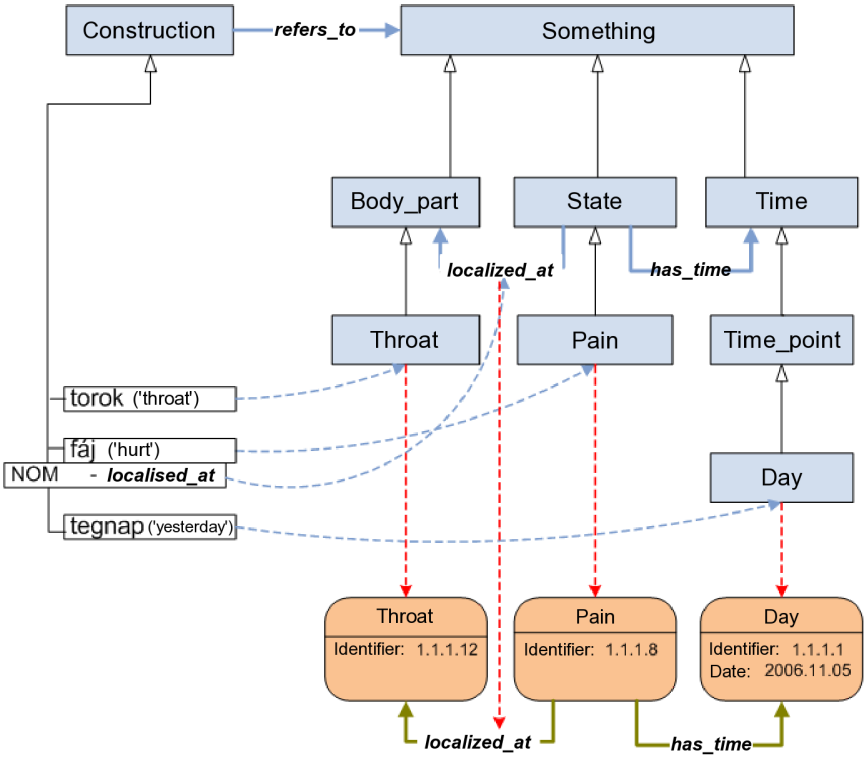

Figure 1: Ontology fragment and semantic representation corresponding to the sentence *Tegnap fájt a torka* '[She/He] had soar throat (her/his throat was hurting) yesterday.' Unlabelled continuous arrows indicate the generic relation. Concepts representing lexemes are connected by the REFERS_TO relation with the corresponding concepts of the world model. The meaning representation is given by the instances shown at the bottom.

3 The linguistic knowledge base and its connection with the world model

In accordance with the MEO model [14], our knowledge base consists of a conceptual layer (the world model) and a linguistic layer that contains the linguistic elements referring to the concepts. Presently, the linguistic component is no more than a lexicon, which lists lexemes (words, affixes, idioms) and their disambiguated versions. The meaning of lexemes are given by restrictions on the REFERS_TO relation (see Figure 1). In addition, a group of lexemes and morphological marks that figure in case-frames also refer directly to concepts or relations. During semantic analysis we create instances of the concepts that correspond to lexemes in the text, and connect them with relations that (might) hold between them according to the

world model. The generated instances can also have data type properties.

Of course, the real problem in meaning representation is not the satisfactory representation of word meanings, but the representation of the syntactic connections that organise words and morphemes into a meaningful sentence—that is, the representation of the head–dependent relation. This can be done in two ways, but the problem of ambiguity has to be faced in both cases.

1. **Representing the case-frame.** A head–dependent relation that is indicated by the grammatical case of the dependent holds between word instances—these syntactic relations correspond to thematic roles in the world model. The correspondence is not universal: for instance, the nominative usually corresponds to the agent or patient role, but in fact it can represent almost any role, e.g. in medical texts it frequently stands for location. Consequently, the mapping between the lexicon and the world model has to indicate the connection between grammatical cases and thematic roles on a case by case basis for each verb: we do this by introducing a relation for each grammatical case (e.g. the relation NOM for the nominative), which determines for every verb or verb phrase the thematic role that corresponds to the case in question. For instance, in the case of *fáj* 'hurt,' the nominative case corresponds to the location role of the PAIN concept, therefore the NOM relation will connect this word with the location thematic role.

2. **Adjuncts.** Most of the affixes that indicate adjuncts in Hungarian can be taken as corresponding to relations of the world model: e.g. the *-ban* 'in' affix can stand for location or time relations.

Naturally, there are expressions to which no referent can be connected in the world model, e.g. articles, negative particles etc. These expressions have to be handled by special rules attached to the syntactic analysis (see Section 5.4).

4 The medical record generating program

The purpose of the software prototype that we have developed is to convert Hungarian texts about medical encounters to a unified medical record representation format. The generated medical record has to contain the patient's identification information, the circumstances of her medical encounters, the reported complaints and symptoms, their properties, the time span of their presence, and possibly other pieces of relevant information.

It is impossible to determine the meaning of a text without relying on syntactic and morphological regularities. Nonetheless, a precise grammatical (especially syntactic) analysis also requires semantic information. In this respect, the ideal solution would be to carry out the morphological, syntactic and semantic analysis simultaneously. Analysers working along these lines already exist for the Hungarian language [1], and our long-term plans also include the implementation of this method. Until then, we consider morphological and syntactic analysis as provided by a preprocessor.

Grammatical preprocessing is performed by MorphoLogic Ltd's morphological and syntactic analyser [10]. The syntactic analysis produces a parse-tree in XML format. The analysis is head-driven: whenever it is possible, larger units, groups, and phrases refer to the terminal element that can be regarded as their head. Relevant morphological characteristics of terminal elements are also indicated, which helps in detecting the semantic connections to a great extent.

Text normalisation is especially important in the case of the medical documents to be processed in the project. We have to handle foreign (typically Latin) words that are characteristic of the subject, abbreviations and their various versions, numbers, and the noticeably frequent mistakes and typos resulting from fast note-taking. Currently, the system works with normalised input. In the next phase we also plan to implement text normalisation in cooperation with MorphoLogic.

To summarise, our software system consists of a grammatical preprocessor, a semantic analyser that generates the semantic representation and a medical record generator that collects those pieces of information from the semantic representation that correspond to fields on the medical record. The grammatical preprocessor is endowed with its own dictionary and grammatical database, while the database of the semantic analyser is the ontology discussed above. The ontology is stored in OWL format, which is a standard, description logic-based ontology language maintained by the W3C consortium [6]. We edit the ontology with the Protégé ontology editor, and our Java code interacts with it using the Jena Semantic Web Framework.

5 Our results

Representing meanings in an ontology raises many problems. Some of these can be solved by adequate design and suitable use of the ontology, but in other cases we have to go beyond the limitations imposed by the ontology and handle the difficulty by external means. We hope that using the previously mentioned lexicalist grammar it will become possible to deal with all of these problems in a uniform way that is internal to the ontology.

5.1 Disambiguation, anaphora resolution

One of the most frequently encountered challenges of free-form text analysis is ambiguity, which appears at several levels of the system. The result of the grammatical analysis can be ambiguous at any of the lexical, morphological and syntactic levels. There can be more than one constructions in the ontology that correspond to a single lexeme, and these constructions in turn will refer to different concepts. Also, in many cases several relations can hold between two concepts.

The presence of anaphoras can also be regarded as a kind of ambiguity, since we usually have to choose from more than one referent candidates. In such cases we look for already processed referents having the properties that are required by the anaphora.

In all of these cases, the basis of disambiguation is the completeness of the competing representations, in the sense of the amount of information they extract from the text under analysis. We calculate this quantity by measuring the specificity of relations in the representation in question: a representation is considered more complete than an other if it contains a larger number of more specific relations.

5.2 Unknown words

However large ontology we build, the text to be analysed will necessarily contain unknown words. We could simply omit these from the semantic representation, but it can easily happen that precisely an unknown word holds together certain parts of the meaning that would otherwise fall apart, since the syntactic structure of the sentence might unambiguously determine the unknown expression's role. Accordingly, we create instances of the THING top category to represent unknown words, and have also introduced the relation DUMMY_RELATION for representing unknown relationships.

5.3 Time and cardinality

Texts frequently do not use tense or do not use only tense to indicate the time of an event or state. In other cases, they contain indexical time adverbs, e.g. *most* 'now' or *két napja* 'for two days.'

Nonetheless, in order to be able to generate the medical record, we have to determine the time span of the complaints as precisely as possible. Consequently, the algorithm contains a module that tries to calculate the time of every complaint and symptom on the basis of the document's time of creation. Our method of representing time in the ontology squares well with this task (see Section 1 and Figure 2).

Cardinalities also present a challenge to ontology designers. Considering the purposes of the project, we opted for broadening the extensions of concepts in order to cover not only single individuals having a certain property, but *sets* of such individuals as well. The fact that an instance is a set is represented by the presence of information about its elements or cardinality. We have distinguished numerical CARDINALITY, which can be expressed by a number, from QUALITATIVE_CARDINALITY (e.g. many, few—see Figure 3).

5.4 External methods

Natural language texts contain many phrases that instead of referring to elements of the world model, modify the meanings of other expressions or their relations to each other. To handle these expressions, we created a dictionary of function words, which contains specific instructions, written in a simple syntax, about the treatment of each listed word.

In the simplest cases, there is nothing to do (e.g. conjuncts like *de* 'but' etc.). In these situations the syntactic analysis already contains the information (e.g.

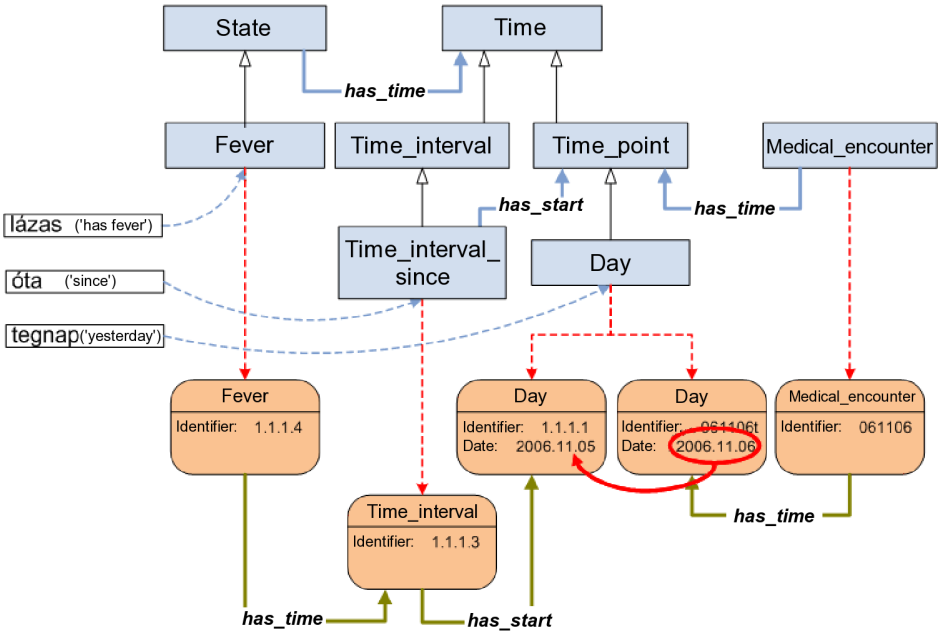

Figure 2: Ontology fragment and semantic representation corresponding to the sentence *Tegnap óta lázas* '[She/He] has had fever since yesterday.' The expression *tegnap óta* 'since yesterday' is represented by a time interval that started on the day preceding the day of the medical encounter when the complaints were recorded.

coordinated clauses in the case of conjuncts), or we do not want to represent the information conveyed by the expression (e.g. the subjective element in the case of *csak* 'only'). Negative particles and affixes are marked during preprocessing, and truth values are determined on this basis. Similarly, we extend dependents connected to lists or coordinated structures to all of the relevant elements already in the preprocessing phase.

There are expressions with special meanings that require complex handling of the sentence in question. For instance, the phrase *egyéb panasza nincsen* 'does not have other complaints' means that the set of complaints (whose elements are all problems and symptoms that were previously mentioned in the text) is closed in the sense that no new element can be added to it.

6 Related work

Although it is still widespread to consider ontologies as consisting of linguistic expressions, a growing number of research projects rely on the distinction made here between ontology, understood as a system of concepts, and expressions that *refer to* these concepts. The OntoWordNet project [5], for instance, aims at working out

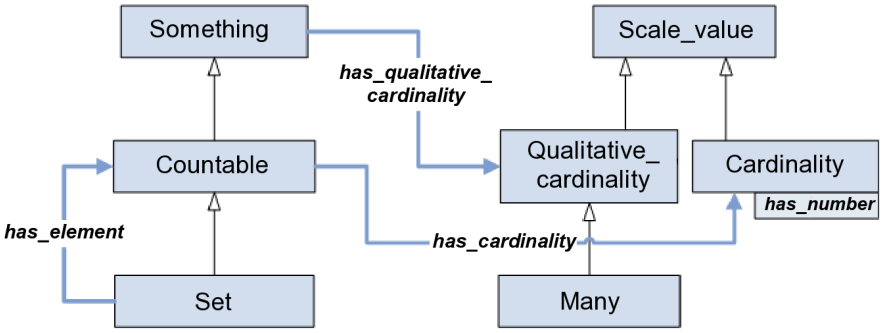

Figure 3: Representation of sets and cardinalities in the ontology.

the connections between the DOLCE foundational ontology [9] and the WordNet lexical database [4], and several projects use ontologies for improving on text search results ([8], [11]), and text generation.

Of the many ontology-based approaches to natural language processing, the closest to our work is the "ontological semantics" presented in [12], although the two projects have been developed independently. Both approaches generate the textual meaning in ontologies, and separate the lexicon from the world model. In contrast to our implementation, which is based on the well established OWL ontology language, ontological semantics implementations use a special, non-standard formalism (so called TMRs) for semantic representation.

An important, relatively early contribution to the semantics of the Hungarian language in the context of natural language processing is also worth mentioning: [3] presents a dependency grammar-inspired system of semantic representation for Hungarian texts, where, similarly to our approach, the semantic links between head verbs and their dependents play a central role.

Even though we do not aim at giving a general semantic theory for natural languages, our work is related to a number of influential approaches in this area as well.

First of all, our ontology based semantic representations share many features of the representations used in Discourse Representation Theory [7]. Both approaches build a partial formal representation of the situation that is described in the represented text by introducing objects for referring expressions with the properties and relations that are explicitly mentioned. The fundamental difference is that in contrast to DRs, our representations are not separated from the general representation of the domain described by the text, but are conceived as extensions of it. This makes it easy to model the dependence of semantic content on conceptual and factual background knowledge.

The idea that the content of a linguistic unit is to be identified with its potential to extend our knowledge of the world is one of the central insights of possible world semantics [13]. One of the main differences here is that the possible world approach does not work with *syntactic* representations: instead, semantic content is identified with classes of possibilities, or mappings between classes of possibilities.

Finally, cognitive linguistics [2] can also be mentioned as a related approach: a specific syntactic construction *together with* its semantic representation in the ontology can be regarded as a linguistically expressed conceptualisation of a situation. In contrast to cognitive linguistics, however, we distinguish the syntactic and semantic layers of this conceptualisation.

7 The future

The medical record generating program we have presented answered many questions, but from the point of view of our long-term plans, it is only a demo having several serious limitations. Further development has to undertake the following tasks.

The architecture of the system is untenable in the long run: the separation of semantic and syntactic analysis is against the philosophy of the method. Therefore, our most important research goal is to work out a grammar with the help of which the syntactic and semantic analysis could be carried out simultaneously—in this way the two processes could cooperate and help each other, e.g. when trying to resolve ambiguities. The grammar in question will be, most probably, a lexicalist grammar. In this case, we will be able to use descriptively the rules that we currently build in the semantic analysis procedurally.

The limitations of the Protégé ontology editor forced us into certain artificial solutions that made the structure of the ontology slightly complicated. We will be able to use a much more natural ontology when the MEO ontology editor will be ready for use [14].

A real life application requires a very extensive ontology, the creation of which would take several years. Therefore we are trying to find a solution that, at least in the case of certain applications, would not require a complete ontology—for instance, the ontology could be built at the time when the program is in use. The most important such application would be a semantic search engine that would produce results based on the meaning of search expressions.

Although the prototype presented here shows the usability of our method, several questions can be raised regarding the analysis of more complicated texts (e.g. the problem of universally or existentially quantified sentences)—we intend to further develop the method and make it capable of coping with these, presently problematic cases as well.

References

[1] Alberti, G., Balogh, K., Kleiber, J., and Viszket, A. A totális lexikalizmus elve és a GASG nyelvtan-modell. In Maleczki, M., editor, *A mai magyar nyelv leírásának újabb módszerei V.*, pages 193–218. Szeged, 2002.

[2] Croft, W. and Cruse, D. A. *Cognitive Linguistics.* Cambridge University Press, 2004.

[3] Farkas, E. and Naszódi, M. Magyar nyelvű mondatok elemzése természetes nyelvű interfész céljából. Technical report, SZTAKI, 1990.

[4] Fellbaum, C., editor. *WordNet: An Electronic Lexical Database.* The MIT Press, 1998.

[5] Gangemi, A., Navigli, R., and Velardi, P. The OntoWordNet Project: extension and axiomatisation of conceptual relations in WordNet. In *International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE 2003)*, Catania, 2003. http://www.loa-cnr.it/Papers/ODBASE-WORDNET.pdf.

[6] Horrocks, I., Patel-Schneider, P. F., and Hayes, P. OWL web ontology language semantics and abstract syntax. W3C recommendation, W3C, 2004. http://www.w3.org/TR/2004/REC-owl-semantics-20040210/.

[7] Kamp, H. A theory of truth and semantic representation. In *Formal Methods in the Study of Language, part 1.* Stichting Matematisch Centrum, Amsterdam, 1981.

[8] Kiryakov, A., Popov, B., Terziev, I., Manov, D., and Ognyanoff, D. Semantic annotation, indexing and retrieval. *Journal of Web Semantics*, (2), 2005.

[9] Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., and Schneider, L. The WonderWeb Library of Foundational Ontologies: Preliminary Report. http://www.loa-cnr.it/Papers/DOLCE2.1-FOL.pdf, 2003.

[10] Merényi, Cs. and Tihanyi, L. A MetaMorpho fordítóprogram projekt 2006-ban. In *MSZNY 2006*, pages 169–179, Szeged, 2006.

[11] Nagypál, G. Improving information retrieval effectiveness by using domain knowledge stored in ontologies. In *OTM Workshops 2005, LNCS 3762*, pages 780–789. Springer, 2005.

[12] Nirenburg, S. and Raskin, V. *Ontological Semantics.* The MIT Press, 2004.

[13] Stalnaker, R. Assertion. In *Syntax and Semantics 9.* Academic Press, New York, 1978.

[14] Szakadát, I., Szőts, M., and Gyepesi, Gy. MEO – ontology infrastructure. In Magyar, G., Knapp, G., Wojtkowski, W., Wojtkowski, G., Zupancic, J., and Wrycza, S., editors, *Advances in Information Systems Development: New Methods and Practice for the Networked Society, Proceedings Information Systems Development*. Springer, In press.

[15] Szőts, M. and Lévay, Á. Szerepfogalmak az ontológiában – az OntoClean metodológia továbbfejlesztése. In *MSZNY 2005*, Szeged, 2005.