Model reprezentace prostorových vztahů

Michal Vavrečka

Abstract:
In the presented paper we analyze the extended version of the model for perceptual grounding of the spatial prepositions. The basic model (Vavrečka, 2007) was able to categorize static spatial relations and create multimodal representations. The extended version is able to create static multimodal representations based on the growing neural networks (GWR) and to categorize dynamic spatial scenes adopting the unsupervised RecSOM architecture. The results are compared to the similar model for the symbol grounding.

Keywords:
symbol grounding, multimodal representations, spatial cognition, GWR, RecSOM

1. Úvod

V oblasti aplikované umělé inteligence narážíme při tvorbě systému na problém, který se týká jejich schopnosti interagovat s okolním prostředím. V posledních desetiletích došlo k rozvoji vtělené robotiky, která se snaží umělé systémy propojit s prostředím pomocí dostatečné senzorické výbavy. Problémy nastávají při zpracování dat ze senzorů. Stále se nedaří odhalit mechanismy vedoucí k tvorbě interní reprezentace, dostačující pro následnou rekonstrukci vjemu či jeho části bez přímé přítomnosti objektu. Současné výzkumy jsou zaměřené na hledání kompromisu mezi mohutností a efektivností reprezentace. Systémy vytvářející reprezentace ze senzorických vstupů dokáží překonat obtíže související s čistě symbolickým způsobem reprezentace, tedy vyřešit problém ukotvení symbolů (Harnad, 1990).

V současné době je pozornost výzkumníků zaměřena na první fázi tvorby senzorických reprezentací, což je spojeno s výzkumem procesů kategorizace. Cílem je tvorba tříd a hierarchií tříd, sloužících umělému systému k optimální reprezentaci okolního prostředí. Při zpracování senzorických informací je nejobtížnější jejich převod do reprezentační úrovně schopné zachytit konstantní vlastnosti a znaky,

tzn. vytvořit konceptuální úroveň schopnou reprezentovat význam a následně tuto úroveň propojit se symbolickou vrstvou.

Jedním z možných řešení je použití multimodálních reprezentací. Samotný pojem multimodální znamená, že vznikají na základě integrace informací ze senzorických modalit (také propriorecepce a introrecepce). Integrace modalit se objevuje již u Damasia (1989), který hovoří o zónách konvergence, které integrují informace ze senzomotorických map, čímž dochází ke společné reprezentaci. Konvergenční zóny pomocí hierarchických množin asociačních oblastí integrují specifické modality. Jedná se o tvorbu mentálních reprezentací napříč modalitami (Barsalou, 2003). V případě prezentovaného modelu se jedná o integraci auditivních a vizuálních modalit. Jelikož auditivní informace má vlastnosti symbolické úrovně (je arbitrární a jednotlivé slova jsou jednoznačně identifikovatelná) a vizuální informace je naopak perceptuální (obsahuje informaci o externím prostředí a je velmi neurčitá (*fuzzy*)), vznikne jejich integrací multimodální reprezentace, ve které jsou spojeny vlastnosti obou.

Model tedy dokáže vytvořit primární reprezentace jednotlivých senzorických vstupů a následně vytvořit společnou reprezentaci. Z hlediska teorie zpracování informace se jedná o výpočetně náročné procesy, přičemž je obtížné určit algoritmickou posloupnost zpracování. Alternativou je použití metody modelování pomocí učících se sítí, které umožní umělému systému tvorbu reprezentací na základě interakce s prostředím. Při adekvátním nastavení počtu neuronů a metod učení můžeme docílit reprezentační úrovně blízké lidským schopnostem.

V rovině teorie je realizace modelu založena na principech kognitivní sémantiky. V případě logiky nepotřebujeme symbolická úroveň ukotvit pomocí perceptuálně získaných konceptů, jelikož význam je založen pouze na koherenci jednotlivých výroku v systému. Význam je definován redukcionisticky. Kognitivní sémantika akceptuje principy logického vyplývání, ale poukazuje na nutnost reprezentace významu také pomocí neverbálních (obecně nesymbolických) mechanismů. Cílem kognitivní sémantiky je nalézt mechanismus převodu perceptů do konceptuální úrovně, která slouží k definici významu. Model založený na integraci senzorických modalit do společné multimodální vrstvy je možnou realizací tohoto cíle, jelikož propojuje konceptuální a symbolickou úroveň.

2. Statický model založený na rostoucích sítích

Ukázkou systému, který reprezentuje okolní prostředí pomocí integrace více senzorických modalit a poukazuje na kontinuální přechod mezi rovinou perceptuální a symbolickou, je navržený model. Dokáže reprezentovat prostorové vztahy

pomocí multimodálních reprezentací, tvořených z výstupů vizuální a auditivní primární reprezentace.

Vnímání prostoru jsem zvolil z několika důvodů. Z neuroanatomického hlediska je informace o poloze a vlastnostech objektů zpracovávána odděleně (Ungerleider, Mishkin, 1982), přičemž rozpoznávání polohy objektu je z hlediska jeho následného zpracování fundamentální. Proto je možné tuto oblast zkoumat samostatně a následně ji integrovat s rovinou rozpoznávání objektu.

Model je zároveň možné analyzovat v kontextu výzkumu ukotvení symbolů, tedy způsobu tvorby konceptuální úrovně a její propojení s úrovní symbolickou. Jelikož existuje mnoho výzkumů či modelů prostorových vztahů, máme zajištěné podmínky pro komparaci navrhovaného řešení. Mezi nejznámější patří Regierův model z roku 1996. Použil síť s více moduly, ve které dochází k ukotvení prostorových vztahů na základě učení s učitelem. V předchozí verzi modelu jsem použil učení bez učitele a dosáhl obdobných výsledků pro statické objekty (Vavrečka, 2007). Jelikož se chci přiblížit psychologické plausibilitě, rozhodl jsem se rozšířit stávající model o rostoucí sítě. Jedná se o architekturu umožňující automatické přidávání či ubírání neuronů pomocí speciálních mechanismů učení na základě vstupních dat. Regierův model obsahoval apriorně daný a fixní počet kategorií (tedy neuronů) pro označení prostorových vztahů, což je v rozporu s *učitelským přístupem* (Ziemke, 1999), jelikož struktura interní reprezentace je predefinována designérem, a nereflektuje vstupní data (např. změnu prostředí). Systém používající rostoucí sítě je naopak schopen vytvářet interní reprezentace flexibilně. V praxi to znamená, že systém se dokáže naučit prostorové označení libovolného jazyka. Pokud nejsou výrazy jednoho jazyka používány (změna prostředí), systém se naučí reagovat na jiný jazyk, tzn. přestane používat stávající neurony nebo je přeučí.

Výsledku je dosaženo také pomocí odlišného návrhu architektury modelu. Oproti jiným řešením (Regier, 1996; Kuipers, 2007) přijímá systém jazyková označení jako auditivní vstup, který je integrován s vizuální informací.

Každý auditivní a vizuální vstup je dvourozměrný vektor, který reprezentuje danou modalitu. V případě vizuálního vstupu reprezentují dimenze vektoru osu X a Y, pomocí které je určena poloha objektu v prostoru. Systém by mohl mít na vstupu umělou sítnici, na kterou by přicházely obraz z prostředí, ale pro simulaci je postačující vektorový vstup, jelikož nedochází k redukci během následného zpracování. Auditivní vstup je oproti jiným řešením také dvourozměrný. Jednotlivé dimenze zachycují fonetické vlastnosti slov, a slouží k vytvoření psychologicky plausibilní primární reprezentaci, ve které jsou si zvukově podobná slova blízká. Jako zvukové dimenze byla zvolena délka slova pro osu X a frekvenční křivka pro osu Y. Dimenzí by mohlo být více, ale pro účely modelu a lepší přehlednost postačují pouze dvě.

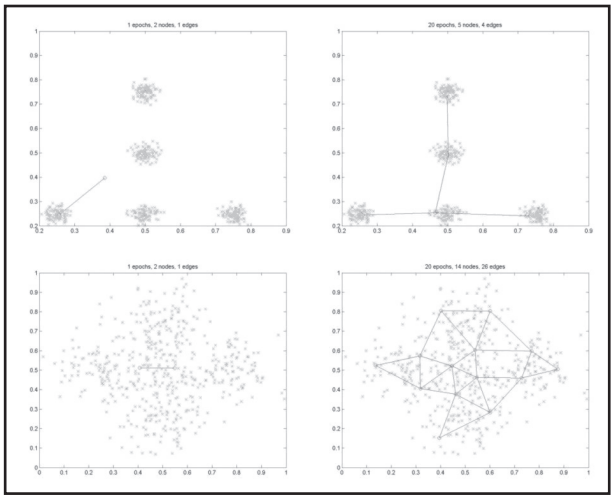
Zpracování auditivního vstupu umožňuje tvorbu jednoznačných primárních kategorií (jazyková označení), které jsou kombinovány s vizuální primární reprezentací (neurony citlivé na polohu v prostoru). Abstraktní rovina ve formě společné multimodální reprezentace je tvořena procesy učení ze dvou odlišných primárních reprezentací. Oproti klasickému přístupu, ve kterém je abstraktní symbolická rovina fixní, jelikož je apriorně definována designérem, v tomto modelu můžeme multimodální reprezentaci měnit na základě učení. Rozdíl je také v sémantice. V čistě symbolických systémech je významem externě daná pravdivostní hodnota. *V navrženém modelu je "významem" propojení auditivních a vizuálních kategorií, které jsou spolu asociovány na základě učení a tvoří společnou úroveň, integrující vlastnosti obou*.

Oproti předchozímu modelu (Vavrečka, 2007) je společná multimodální vrstva tvořena sítí typu GWR – Growing when required (Marsland, 2005). V prvotní fázi učení obsahuje výstupní vrstva pouze dva neurony (viz obr. 39). Algoritmus učení však umožňuje přidávání nových neuronů, pokud síť nedokáže kategorizovat s dostatečnou přesností. Klíčovým parametrem je nastavení citlivosti (insertion threshold), který určuje konečný počet neuronů. Závislost změny citlivosti kategorizace na počet neuronů je zachycen v tab. 5.

Insertion threshold	*Počet neuronů*	*Insertion threshold*	*Počet neuronů*	*Insertion threshold*	*Počet neuronů*
0.00001	3	0.00011	10	0.00021	11
0.00002	3	0.00012	11	0.00022	11
0.00003	3	0.00013	11	0.00023	13
0.00004	5	0.00014	12	0.00024	12
0.00005	5	0.00015	11	0.00025	13
0.00006	5	0.00016	11	0.00026	12
0.00007	8	0.00017	11	0.00027	13
0.00008	8	0.00018	11	0.00028	12
0.00009	9	0.00019	11	0.00029	12
0.0001	10	0.0002	11	0.0003	12

Tab. 5 – Závislost počtu přidaných neuronů na citlivosti algoritmu učení (insertion threshold). Síť byla trénována 20 epoch, learning rate 0.3 pro vítězný neuron a 0.01 pro sousední neurony. V rozmezí hodnot 0.00004–0.00006 dokáže síť vytvořit správný počet kategorií pro lokalistickou reprezentaci pěti prostorových označení (vpravo, vlevo, nahoře, dole, uprostřed). Se zvyšující se hodnotou síť vytváří redundantní počet neuronů, čehož lze využít pro distribuované reprezentace.

Z tabulky lze vyčíst, že nastavením parametru je síť schopna odvodit adekvátní počet výstupních kategorií na základě vstupních dat, což je ve shodě s *učitelským principem*. Pomocí změny parametru můžeme navíc modifikovat robustnost reprezentace. V případě, že počet neuronů ve výstupní vrstvě odpovídá počtu prostorových kategorií, získáváme lokalistické reprezentace, které můžeme použít jako vstup např. pro následnou symbolickou část modelu (jeden neuron ve výstupní multimodální vrstvě odpovídá jednomu prostorovému vztahu). Popřípadě změnou parametru a výstupní funkce vytvoříme distribuované reprezentace, tzn. že každá prostorová kategorie je reprezentována pomocí několika neuronů, v jejichž výstupní aktivitě je zakódována prostorová poloha, což vede k větší odolnosti vůči šumu. V další fázi vývoje modelu je možné měnit nastavení parametru pomocí vyšších vrstev sítě, což by umožňovalo síti modifikovat robustnost multimodální reprezentace vzhledem k obtížnosti řešené úlohy, tzn. adaptivní seberegulaci. Podobný princip nacházíme u sítí typu ART (Carpenter, Grossberg, 2003).



Obr. 39 – Ukázka činnosti rostoucích sítí. Na začátku jsou vytvořeny dva náhodné neurony, které jsou trénovány pomocí učení bez učitele. Pokud není kategorizace dostačující, což je řízeno parametrem insertion threshold, síť automaticky přidá nové neurony. Vlevo je zobrazena primární audio (nahoře) a video (dole) reprezentace na začátku učení (pouze 2 neurony), vpravo po 20 epochách (audio vrstva obsahuje 5 neuronů, což odpovídá počtu prostorových označení; video vrstva obsahuje 14 neuronů, které slouží ke kategorizaci vizuálního pole na jednotlivé části). Audio vstupy: osa X – délka slova, osa Y – frekvenční křivka slova; Video vstupy – poloha bodu v 2D prostoru pomocí souřadnic X a Y. Vstupní vektory jsou označeny křížkem, neurony kolečky a jejich vzájemné vazby čárami.

Již po 20 epochách učení dokázala síť vytvořit optimální počet neuronů pro lokalistickou i distribuovanou multimodální reprezentaci. Model byl schopen vytvořit multimodální reprezentace pěti prostorových kategorií. Jelikož vstupy multimodální vrstvy jsou mnohodimenzionální, je pro vetší přehlednost na obrázku zachycena geneze GWR sítí pro primární auditivní a vizuální reprezentace, které jsou tvořené na základě dvourozměrných vstupů.

Schopnost přidávat nové neurony má několik opodstatnění. Jelikož dosud neznáme přesné fungování vizuální dráhy pro zpracování prostoru, snažíme se napodobit její strukturu pomocí sítě, která se sama vytváří, přičemž jsou jí prezentovány stejné podněty jako lidskému mozku. Síť tedy adaptivně reaguje na vnější podněty a přidáváním nových neuronů hledáme jejich optimální množství pro následné testování na různých typech úloh. Zároveň eliminujeme vliv apriorních předpokladů při konstrukci systémů, jelikož model se během své činnosti modifikuje sám.

3. Model dynamických prostorových vztahů

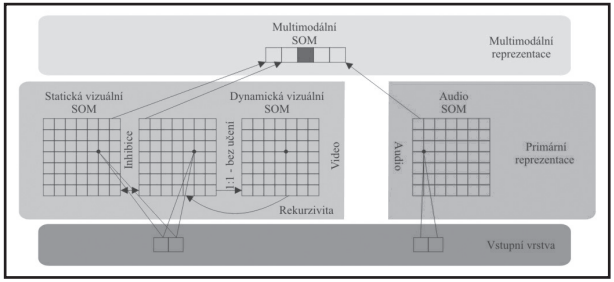
V případě reprezentace dynamických scén je situace komplikovanější, protože musíme brát v potaz temporální informace. Je potřeba nalézt vhodný způsob, jakým reprezentovat průběh pohybu v čase. Jelikož při reprezentaci času roste výpočetní mohutnost, přistupují k této problematice badatelé odlišně. Regier (1996) apriorně označil objekt, u kterého chtěl zjistit prostorový vztah vzhledem k druhému ,a poté zjednodušil průběh pohybu na sekvenci *počátek–cesta–cíl*. Z nich odvozoval typ pohybu, přičemž u položky *cesta* systém detekoval, zda došlo k doteku či průniku sledovaného objektu, pro rozlišení prostorových vztahu „skrz“, „po“ a „pod“.

Siskind ve svém modelu (2001) řeší reprezentaci prostorových vztahů pomocí převodů elementárních označení do symbolické roviny. Označení řetězí formou výroků, se kterými následně pracuje na principech klasické logiky. Systém si nejprve na základě percepce vytvoří sémantická primitiva (na podobném principu je založena Cangelosiho *symbolická krádež*), se kterými je dále manipulováno v symbolické rovině dle pravidel formální logiky. Jeho systém obsahuje modul na rozpoznání elementárních objektů pro identifikaci jejich vzájemného propojení. Vychází z Talmyho „force dynamics“ (1988), kterou modifikuje na prozkoumávání vztahů typu „podpírá“, dotýká se“ a „je propojen“. Jedná se o řešení základních interakcí mezi objekty, vyžadující znalost o jejich tvaru.

V navrženém modelu se zabývám elementárnějšími prostorovými vztahy, jelikož výše zmíněné vyžadují identifikaci okraje objektů a dalších vlastností, tzn. tvorbu specializovaného modulu. V našem případě je pozornost soustředěna pouze na informace o poloze objektu. Systém je daleko univerzálnější, protože většina zmíněných výzkumů je založena na apriorních předpokladech či omezeních týkajících se rozpoznání objektů, což může být zavádějící pro jejich následnou interpretaci.

Co se týká změn ve způsobu zpracování informací, není symbolická úroveň systému apriorně daná, ale učí se jí na základě auditivního vstupu kombinovaného

se vstupem vizuálním. Pro reprezentaci dynamických scén slouží speciální vizuální subsystém, který je založen na rekurzívních sítích typu RecSOM (Voegtlin, 2002). Jeho samostatnost je v souladu z biologickými a psychologickými poznatky. Lidské oko obsahuje buňky citlivé na statickou polohu objektů a také buňky citlivé na pohyb. Sítě RecSOM jsou výhodné právě pro svou schopnost reprezentace pohybu. Oproti klasický sítím SOM (Kohonen, 1989) obsahují kontextovou vrstvu, která dokáže zachytit časové proměny a vhodným způsobem je reprezentovat. Schéma rozšířené verze modelu je vidět na obr. 40.



Obr. 40 – Základní schéma modelu. Audio a video vstup je reprezentován odděleně pomocí SOM sítí. Následně jsou výstupy integrovány v multimodální vrstvě.

Samotná činnost modelu probíhá následovně. Síti jsou prezentovány trénovací vzorky jednotlivých typů dynamických prostorových vztahů (přes, pod, skrz, ven a okolo) ve formě sekvence, která se skládá z deseti dvourozměrných souřadnic popisujících trajektorii pohybu. Po natrénování obsahuje výstupní vrstva neurony citlivé na specifické změny polohy. Celkový pohyb je následně reprezentovány jako sekvence změn a výstupní funkce posílá do multimodální vrstvy sekvenci vítězných neuronů. Informace je zpracována společně s auditivním vstupem (označení prostorových vztahů) a dochází ke tvorbě multimodálních kategorií. Čas je reprezentován jako sekvence výstupní funkce neuronů citlivých na jednotlivé změny pohybů, pomocí které můžeme odlišit jednotlivé typy dynamických prostorových vztahů již v rovině zpracování vizuálního podnětu a vytvořit jejich elementární kategorie. Auditivní informace je pro shodný způsob prezentace zpracovávána obdobně jako u statických prostorových vztahů. Model je rozšířen pouze o rekurentní síť pro reprezentaci pohybu, jinak je identický s modelem předchozím.

Realizace modelu proběhla pomocí programu Matlab s využitím toolboxu Neural Networks. Během testování bylo použito také specializovaných modulů SOMToolbox pro statické vztahy a RecSOMToolbox pro dynamické vztahy. Pro rostoucí sítě byly použity algoritmy GWR od Stephena Marslanda. Většina zmíněných nástrojů byla modifikována pro použití v popsaném modelu a jejich zdrojový kód je zájemcům na vyžádání k dispozici.

Diskuze

Jak vyplývá z předchozího textu, realizovaný model se odlišuje od ostatních výzkumů v několika ohledech. Pozornost je soustředěna pouze na zpracování prostorové polohy objektu bez přesné identifikace jeho tvaru. Důvodů pro takový postup je několik. Regierův model (1996) obsahuje apriorní předpoklad o centrálním a k němu vztaženém objektu. Vytváří systém, ve kterém je centrální objekt umístěn vždy ve středu a relativní objekt zaujímá prostorové pozice. Je tedy vytvořen absolutní systém, který není schopen reflektovat změnu polohy senzoru (oka), tzn. přesunu centrálního objektu na periferii. V našem modelu je nahrazen centrální objekt centrem senzoru a změna polohy se projeví změnou prostorového vztahu vzhledem k pozorovateli. Přesto je model, podobně jako Regierův, založen na absolutním prostorovém systému. Jak již bylo zmíněno v předchozím článku (Vavrečka, 2007), nelze pomocí jediného absolutního systému zajistit identifikaci více než jednoho objektu v prostoru (vzhledem k pozorovatelovu centru senzorického pole), a proto je plánováno rozšíření o relativní prostorové moduly, které by umožňovaly identifikaci polohy více objektů.

Také Siskindův model obsahuje apriorní předpoklady o tvaru objektu, sloužící systému k rozpoznání vztahů typu „podpora", „dotek", „propojení". V současné fázi výzkumu však není možné zajistit univerzálnost reprezentace těchto vztahů, jelikož identifikace objektů není na dostatečné úrovni, abychom zajistili spolehlivost rozpoznání různých tvarů a typů objektů. Většina navrhovaných modelů obsahuje velké množství požadavků na schopnost systému, které lze uskutečnit pouze zjednodušením zadaní a z toho plynoucího zjednodušení subsystému pro rozpoznávání objektů .

Proto realizovaný model tyto prostorové vztahy nezahrnuje. Pozornost je soustředěna pouze na základní prostorové vztahy, které je možné realizovat v jediném absolutním prostorovém systému bez nutnosti identifikace tvaru objektu. Model nedokáže rozlišit vztahy, při kterých dochází k interakci dvou objektů na úrovni jejich vzájemného propojení či nulové vzdálenosti. Na druhou stranu jsou realizované prostorové reprezentace univerzálně použitelné ve všech typech prostředí s libovolnými objekty a libovolným typem jazykových označení. Pro samotný výzkum způsobu ukotvení symbolů, přesněji řečeno převodu senzorických informací do abstraktní úrovně, je navrhovaný systém dostačující. Rozšíření modelu nám umožní pouze možnost testovat systém na složitějším typu úloh.

V dalších fázích bude model rozšířen o zmíněný relativní prostorový systém, který využívá informace ze stávajícího absolutního modulu a také o schopnost reprezentovat dynamické prostorové vztahy prezentované v libovolném počtu kroků. Hlavním cílem, který souvisí s problematikou ukotvení symbolů, je následné zpracování

v symbolické rovině a tvorba kompozicionality. Model tak získá schopnost reprezentovat okolní prostředí a vytvářet o něm závěry ve formě výroku, jejichž pravdivostní hodnota není do systému vkládána externím pozorovatelem, ale je systémem budována a reprezentována na základě součinnosti jeho senzorických vstupů.[1)]

Literatura:

BARSALOU, L. W. Perceptual symbol systems. *Behavioral and Brain Sciences* 22, 1999, s. 577–609.

CARPENTER, G. A.; GROSSBERG S. Adaptive Resonance Theory. In ARBIB, M. A. (ed.) *The Handbook of Brain Theory and Neural Networks*, 2nd ed., Cambridge, MA: MIT Press, 2003, s. 87–90.

HARNAD The Symbol Grounding Problem. *Physica D*, 1990, s. 335–346.

KOHONEN, T. *Self-organization and associative memory*. New York: Springer, 1989.

KUIPERS, B. An intellectual history of the Spatial Semantic Hierarchy In YEAP, M. J.; YEAP, A. (Wai-Kiang) (eds.) *Robot and Cognitive Approaches to Spatial Mapping To appear*, Verlag: Springer, 2007.

MARSLAND, S. A self-organising network that grows when required. *Neural Networks* 15, 2002, s. 1041–1058.

REGIER, T. *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Cambridge, MA: MIT Press, 1996.

SCHILLS, K. Hybrid architecture for the sensorimotor representation of spatial configurations. *Cogn Process* 7 (Suppl. 1), 2006, s. S90–S92.

SISKIND, J. M. Grounding the Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic. *J. Artif. Intell. Res. (JAIR)* 15, 2001, s. 31–90.

UNGERLEIDER, L. G.; MISHKIN, M. Two cortical visual systems. In INGLE, D. J.; GOODALE, M. A.; MANSFIELD, R. J. W. (eds.) *Analysis of visual behaviour* . Cambridge, MA: MIT Press, 1982, 549–586.

VARELA, F.; THOMPSON, E.; ROSCH E. *The Embodied Mind – Cognitive Science and Human Experience*. Cambridge, MA: MIT Press, Bradford Books, 1991.

VAVREČKA, M. Ukotvení prostorových vztahů. In KELEMEN, J.; KVASNIČKA, V.; TREBATICKÝ, P. (eds.) *Kognicia a umelý život VII.* Bratislava, 2007.

VOEGTLIN, T. Recursive self-organizing maps. *Neur. Netw.* 15/8–9, 2002, s. 979–991.

ZIEMKE, T. Rethinking Grounding. In RIEGLER, A.; PESCHL, M.; VON STEIN, A. (eds.) *Understanding Representation in the Cognitive Sciences*. New York: Plenum Press, 1999, s. 177–190.

1) Chtěl bych poděkovat doc. Igoru Farkašovi z Katedry aplikované informatiky a matematiky Komenského univerzity v Bratislavě za pomoc při realizaci modelu a RNDr. Martinu Takáčovi za inspirativní připomínky.