Reprezentace objektů v prostoru pomocí dvou vizuálních drah

Michal Vavrečka

Biodat, FEL ČVUT
Karlovo náměstí 13, Praha 1
vavrecka@fel.cvut.cz

Abstrakt

V příspěvku se zabývám rozborem modelu, který slouží k reprezentaci objektu v prostoru. Systém je tvořen umělou retinou, na kterou jsou promítány polohy dvou rozličných objektů v prostoru, přičemž je zároveň prezentován fonetický vstup popisující scénu. Systém integruje vstupy z umělého oka a ucha a vytváří multimodální reprezentace. Cílem experimentu bylo porovnat systém pracující s jedním vizuálním subsystémem a systém zpracovávající zvlášť informaci o tvaru a barvě objektu a jeho prostorové poloze. Efektivita těchto systémů byla testována na podnětech se vzrůstající mírou komplexnosti. Architektura systému je založena na sítích typu SOM a RecSOM.

Klíčová slova: reprezentace prostoru, multimodální reprezentace, SOM, RecSOM

1 Úvod

Při návrhu modelu vycházím z principů kognitivní sémantiky, přičemž základní inspiraci nacházím již v Peircově teorii znaku (1931-35). Hovoří o tvorbě znaku (reprezentace), který referuje k externímu objektu pomocí konceptu (interpretant) a arbitrární symbolické vrstvy (representamen). V realizovaném modelu (Vavrečka, 2008) je konceptuální úroveň (koncept) vytvářena na základě perceptuálních vstupů, což je v souladu s perceptuální teorií kognice (Barsalou, 1999) a zajišťuje korespondenci interních reprezentací s externím prostředím. Tím se přístup liší od Saussureovy teorie znaku (1965) a také formální sémantiky v rámci analytické filosofie (Tarski, 1944). Oproti formálnímu symbolickému systému a formální sémantice je tento systém odlišný právě užitím perceptuálně získané konceptuální úrovně.

Realizovaný model (Vavrečka, 2008) je v teoretické úrovni řešením problematiky ukotvení symbolů (Harnad, 1990), která se zabývá způsobem převodu perceptu do konceptuální úrovně a její následné propojení s arbitrárním symbolickým označením. Jeden z možných způsobů převodu, který je inspirován teorií perceptuálního symbolu, je tvorba multimodálních reprezentací, tzn. integrace informací z různých senzorických modalit do společného rámce. Podobnou inspiraci nacházíme v psychologii jako teorii duálního kódování (Paivio, 1986), v rovině kognitivního modelování v podobě Dorffnerova modelu ukotvení (1996). Systém nejprve primárně reprezentuje auditivní a vizuální vstupy a poté je integruje do společné multimodální vrstvy. Následně nás zajímá otázka, který vstup se na strukturaci multimodální vrstvy podílí.

V oblasti psychologie a lingvistiky se totiž setkáváme v souvislosti s propojením symbolické a konceptuální roviny s otázkami, která z těchto úrovní je hierarchicky výše. Jedná se o spory ohledně privilegovanosti lingvistické (symbolické) či nelingvistické (konceptuální) reprezentace (Landau, Hofmann, 2005). Použitím multimodální vrstvy, která tyto dvě úrovně integruje, byly zajištěny podmínky pro analýzu způsobu jejich činnosti. Výsledky fungování předchozího modelu vedly k postulaci principu privilegovanosti jednoznačného vstupu (Vavrečka, 2008), podle kterého jsou reprezentace strukturovány vstupem, který je v danou chvíli jednoznačný, tzn. vytváří ohraničené kategorie.

Jelikož je stávající i navrhovaný model realizován v oblasti vnímání prostoru, je potřeba zmínit detaily, které s touto problematikou souvisí. Vycházíme z Coventryho FGR (Functional Geometric Framework) teorie (2004), která slouží k odvození prostorových vztahů na základě a) geometrických vztahů a b) extra-geometrických vztahů. Jedná se o rozlišení základních kategorií, spolupodílejících se na identifikaci prostorového označení mezi objekty. Geometrické vztahy souvisí s prostorovou polohou objektu. Systém extra-geometrických vztahů doplňuje informace o gravitaci, silách působících mezi objekty, vlivech kontextu a dalších aspektech, které společně s předchozím systémem determinují označení prostorového vztahu. Dohromady tyto části tvoří jednotný funkční geometrický rámec (FGR). Oproti předchozím přístupům (Miller, Johnson-Laird, 1976;Herskovits, 1988) se jedná o integrovaný systém, který zohledňuje všechny vlastnosti, potřebné ke správnému odvození prostorového vztahu.

Jelikož je systém extra-geometrických vztahů komplexní a kontextově závislý, byly v současné fázi modelu zohledněny pouze geometrické vztahy.

2 Model I

2.1 Vstupy

V první fázi se jednalo o tvorbu modelu, který by byl schopen samostatně rozpoznávat polohu a tvar objektu. Přestože biologicky plausibilním řešením je tvorba dvou samostatných vizuálních subsystémů, byly v první verzi modelu oba systémy integrovány do jediné společné vrstvy neuronové sítě. V praxi to znamená, že byla vytvořena dostatečně velká SOM mapa, která dokáže reprezentovat jak změnu prostorových vztahů objektů tak jejich tvar a barvu.
Vizuální a auditivní vstupy byly generovány pomocí vytvořeného programu pro tvorbu scén, který dokáže kombinatoricky pokrýt veškeré možnosti, které jsou modelu prezentovány. V praxi se jednalo o tvorbu dvou až pěti základních objektů (krabice, míč, stůl, bedna, pohár, postel) ve 2-3 barvách, které se vyskytovaly ve 4 různých vzájemných prostorových relacích (nahoru, dolu, vlevo vpravo). Generátor tedy vytvoří scény, ve kterých se jeden objekt vyskytuje uprostřed a druhý objekt je umístěn v jeho okolí.

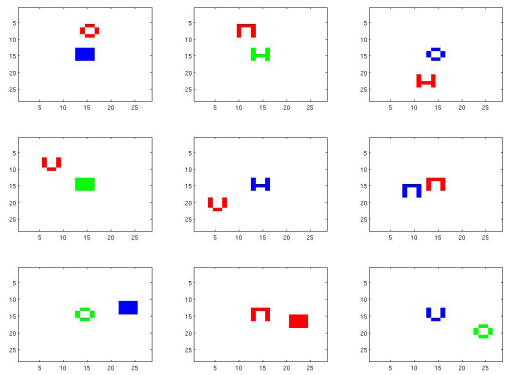

Fig. 1 Ukázky vstupů vizuálního subsystému

Jelikož předchozí model reprezentoval znalosti pouze na úrovni jednotlivých slov, bylo třeba upravit auditivní vrstvu architektury tak, aby dokázala zpracovávat celé věty popisující vztah dvou objektů. Proto byla použita architektura RecSOM, která dokáže zachytit sekvence znaků, podobně jako je tomu při zpracování věty jako řetězce slov. Zpracování vizuálního vstupu bylo podobné jako u předchozího modelu (SOM mapa). Architektura doznala změny ve vstupní vrstvě, jelikož byla u obou modalit doplněna o umělou retinu a také umělý analyzátor slov.

2.2 Vizuální subsystém

Vizuální subsystém byl v novém modelu tvořena pomocí umělé sítnice skládající se z mřížky o velikosti 28x28 neuronů na kterou byla promítány jednotlivé scény. Sítnice je plně propojená s primární (unimodální) vizuální vrstvou, která slouží k reprezentaci jednotlivých scén. Je tvořena SOM mapou, která by měla být schopná se naučit rozlišovat jednotlivé polohy dvou objektů v prostoru a také rozlišovat jednotlivé objekty a barvy. Síť byla trénována po dobu 100 epoch, přičemž learning rate klesá z počáteční hodnoty 0.3 až na 0. Síť měla čtvercový tvar a podobně jako u předchozího modelu (Vavrečka, 2008) jsme zvětšovali velikost mřížky a testovali, jakým způsobem se mění schopnost reprezentace znalostí. Počet neuronů v mřížce vzrůstal od hodnoty 5x5 až do 40x40 neuronů. Protože není u menší velikosti mřížky kombinatoricky možné reprezentovat více druhů objektů, slouží nám pouze k zachycení tvorby reprezentací a pochopení geneze učení. Po natrénování byla síti prezentována sada testovacích vzorů (vytvořena generátorem) a zjištěna efektivita učení. Pro lepší pochopení geneze reprezentací jsme vizualizovali matice vah pro jednotlivé neurony, pomoci nichž jsme schopni určit způsob kategorizace jednotlivých vstupů (Fig.2)

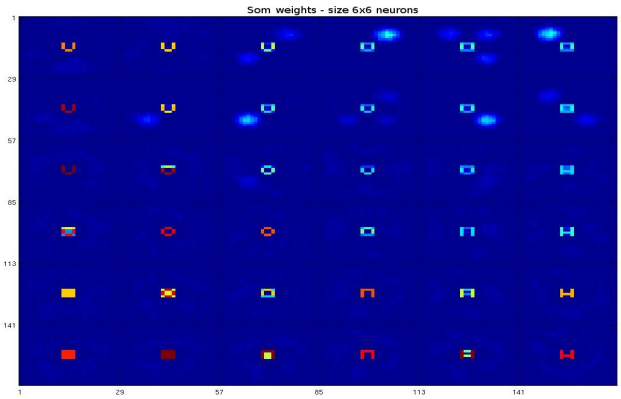

Fig. 2 Při velikosti 6x6 síť už dokáže rozlišit jednotlivé báze, které se nehýbou a také se objevují neurony citlivé na daný prostorový kvadrant.

Se stoupajícím počtem neuronů v mřížce, je síť schopná rozlišovat veškeré objekty báze, ve všech barevných provedeních a také čtyři prostorové oblasti jsou dobře zachytitelné v matici vah. Ani při velikosti mřížky 30x30 však není síť schopná jednoznačně reprezentovat složitější kombinace objektů (3, barvy,5 tvarů, 4 prostorové vztahy). Zkoušeli jsme testovat síť až do velikosti 45x45 neuronů, ale stoupá pouze schopnost rozlišovat prostorové vztahy, ale nikoliv rozpoznávat trajektory. Jelikož je variance trajektoru v prostoru náhodná pro jednotlivé směry, může být tento počet příliš malý pro schopnost naučit se tvar objektu. Také způsob překrývání u jednotlivých objektů je pro systém složitým problémem. Oproti podobným simulacím, kdy je objekt v prostoru prezentován v diskrétních polohách jsme chtěli otestovat variantu s překrývajícími se polohami trajektoru. Jedná se však již o úlohu, kterou síť nedokáže zvládnout.

Z uvedených simulací vyplývá, že primární vizuální vrstvu je potřeba skládat ze samostatných subsystémů pro zpracování identifikace objektu a identifikace polohy, podobně jako je tomu i biologických systémů. Podrobnější analýzou navrhovaného systému bychom došli k závěru, že mapování polohy i tvaru dvou objektů je kombinatoricky příliš náročné a je potřeba úlohu dekomponovat na dílčí cíle (viz. Model II).

2.3. Auditivní subsystém

Auditivní vstup je plně propojen s primární RecSOM vrstvou, která slouží k reprezentaci jednotlivých slov a jejich pořadí. Věta je prezentována systému po jednotlivých slovech, přičemž kontextová vrstva v síti RecSOM umožňuje zachytit vztahy mezi předchozím a následujícím slovem. Cílem simulace je vytvořit auditivní vrstvu schopnou reprezentovat kontext. Výstup auditivní vrstvy je napojen na následující multimodální vrstvu, kde dochází k integraci s výstupy z primární vizuální vrstvy. Trénování sítě probíhalo pomocí výše zmíněných vstupů, které byly přes vstupní vrstvu reprezentovány v RecSOM síti. Poté jsme testovali a zjišťovali od jaké velikosti je možné odlišovat polohu slova ve větě. Měnili jsme velikost RecSOM mřížky a při testování zaznamenali efektivitu. Oproti vizuální vrstvě je auditivní mřížka méně náročná, jelikož je věta popisující daný vztah reprezentována sekvencí, což činí úlohu méně kombinatoricky náročnou. Navíc auditivní vrstva obsahuje kontextovou vrstvu, která dokáže uchovávat temporální informaci (Fig.3)

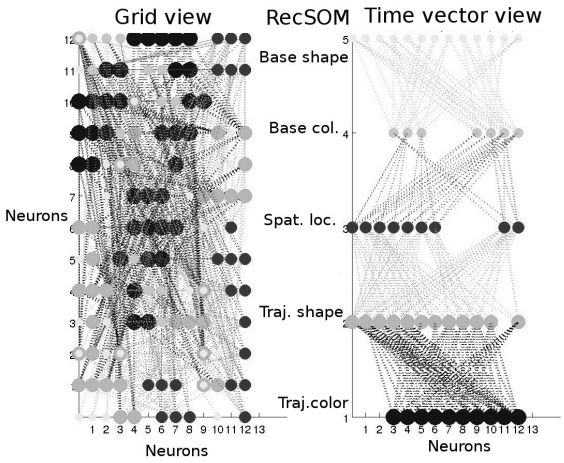

Fig. 3 Vizualizace RecSOM sítě pomocí časových trajektorů (pohled shora a časový průběh slov ve větě). Síť o velikosti 12x12 dokáže rozlišovat barvy (časová vrstva 1 a 4 odspodu) ale při označení trajektoru a báze dochází k „binding“ problému (vrstva 2 a 5).

2.4. Multimodální vrstva

Multimodální vrstva je tvořena také pomocí SOM mřížky, která se musí naučit propojení z jednotlivých modalit. Její velikost je nastavená tak, aby byla schopná lokalisticky reprezentovat všechny možné kombinací objektů i v případě nejsložitější úlohy (3 barvy, 5 objektů, 4 prostorové vztahy). Tvoří ji 29x29 neuronů (840 možných kombinací). Vstupy přicházejí z primárních vrstev a výstupem multimodální vrstvy je vítězný neuron, který reprezentuje danou prostorovou konfiguraci objektů. Výstupní funkce je navržena tak, aby bylo možné změřit přesnost kategorizace jednotlivých vstupů , nejedná se tedy o distribuovanou reprezentaci.

Síť byla trénována pomocí výstupů z auditivní a vizuální vrstvy, přičemž dle předchozího modelu byla zachována symetrie velikosti jejich mřížek, aby nedocházelo k distorzi během učení. Abychom zachytili genezi vývoje systému, použili jsme k učení zvětšujících se unimodálních vrstev, přičemž multimodální vrstva byla konstantní. Výsledky potvrdily problémy v primární vizuální vrstvě s rozpoznáváním trajektoru, což nás vedlo k návrhu řešení, založeného na samostatném zpracování polohy a tvaru objektu.

3 Model II

Abychom potvrdili jako zdroj nepřesnosti vizuální vrstvu,

provedli jsme simulaci, ve které byla rozšířena vizuální unimodální vrstva. Systém obsahoval kromě zmíněné SOM mapy pro reprezentaci prostorových vztahů, také druhou SOM mapu simulující systém pro rozpoznávání objektů. Vstupní část sítě simulovala foveu, na kterou jsou promítány pouze objekty v centru pozornosti (trajektor a následně báze). Její rozměr byl 8x4 neurony a výstupy byly převáděny na unimodální reprezentaci objektu, tvořené čtvercovou mřížkou. Celý systém je identický s modelem I, jediné rozšíření tvoří *what* systém. Abychom zjistili jaké faktory se podílejí na nepřesnosti v unimodální vrstvě, porovnávali jsme oba modely pomocí identických vstupů. Pro detekci rozdílů jednotlivých modelů při vzrůstající komplexitě, byly vygenerovány trénovací a testovací data, sestávající z rozdílného počtu barev a objektů. Nejjednodušší sada obsahovala 2 objekty, 2 barvy, 4 prostorové vztahy a 100 příkladů pro jednu kombinaci (celkem 6400 kombinací). Nejsložitější varianta 3 barvy, 4 objekty, 4 prostorové vztahy a 50 příkladů pro jednu kombinaci (28800 trénovacích vzorků). Jelikož rozdíly pro nejjednodušší variantu nebyly veliké a obě typy architektur se naučily reprezentovat polohy vlastnosti bez chyby, uvádíme výsledky pouze dvou komplexnějších variant (Fig.4-5).

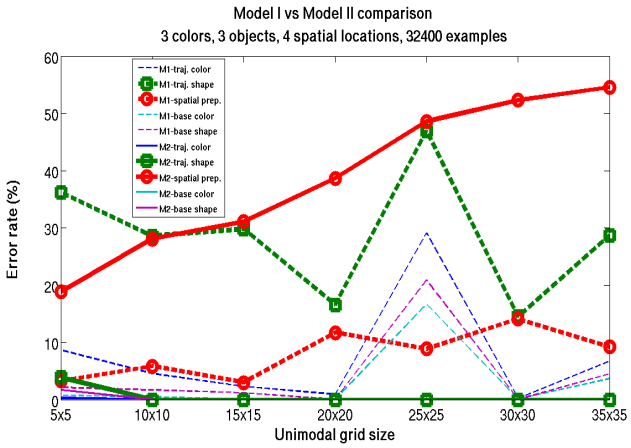

Fig. 4 Porovnání modelu 1 (čárkovaně) a modelu 2 (plná čára) z hlediska chyby v multimodální vrstvě pro reprezentaci jednotlivých vlastností objektů (2 barvy, 3 objekty, 4 prostorové vztahy, 32400 příkladů). Na ose X je zachycena stoupající velikost unimodálních mřížek, na ose Y chyba po natrénování.

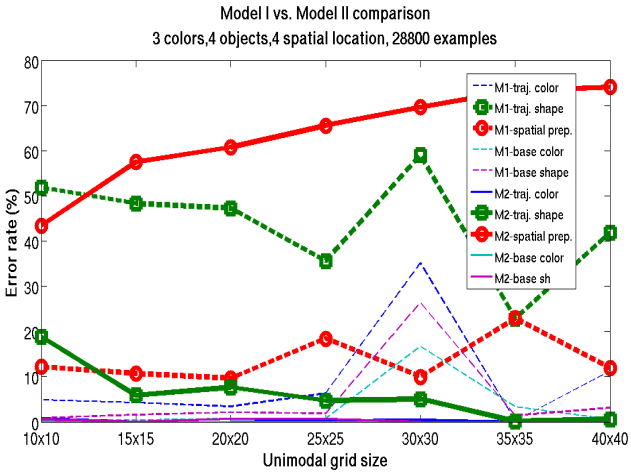

Fig. 5 Porovnání modelu 1 (čárkovaně) a modelu 2 (plná čára) z hlediska chyby v multimodální vrstvě pro reprezentaci jednotlivých vlastností objektů (3 barvy, 4 objekty, 4 prostorové vztahy, 28800 příkladů). Na ose X je zachycena stoupající velikost unimodálních mřížek, na ose Y chyba po natrénování.

3 Diskuze

Jak vyplývá z tabulky, jsou výsledky porovnání obou typů architektur nejednoznačné. Společné znaky nacházíme v případě reprezentace barev trajektoru a báze a tvaru trajektoru. Obě architektury jsou schopny rozlišovat tyto vlastnosti z chybou okolo 5 %, přičemž architektura se samostatnou reprezentací polohy a tvaru objektu dosahuje lepších výsledků. Schopnost správné reprezentace báze je dána její fixní polohou, jelikož systém nemusí pracovat s prostorovou variabilitou a stačí pouze odlišit jednotlivé tvary.

Z hlediska auditivní reprezentace je situace obdobná, protože jednotlivá označení nedisponují velkou variabilitou a systém dokáže v obou případech reprezentovat věty již od velikosti mřížky 15x15. Během testování RecSOM verze se zapnutým nebo vypnutým resetováním kontextové vrstvy po každé větě, jsme získali lepší výsledky při vypnutém resetování, což je biologicky plausibilnější přístup.

Největší rozdíly mezi jednotlivými architekturami nacházíme v oblasti reprezentace tvaru trajektoru a prostorového označení. Pro model obsahující *what* a *where* systém v jedné SOM mapě se objevuje největší chyba v identifikaci tvaru trajektoru (v grafu symbol čtverečku). Je to způsobeno jeho variabilitou a také neschopnosti jediné SOM mapy reprezentovat tvar, barvu i polohu dvou objektů. V případě modelu II znamenalo rozšíření systému o samostatný *what* systém snížení chyby v průměru o 30 procent na hodnoty kolem 0.

Systém dostává z *what* systému jednoznačnou informaci o tvaru trajektoru, takže výsledná multimodální reprezentace je přesnější.
Rozšířená verze architektury ale zároveň způsobuje zvýšení chyby prostorového označení (v grafu symbol kolečka). Jak je patrné, vzrostla chyba ve správném označení o 40-50 %. Příčiny vzniku této chyby můžeme hledat ve zmíněném *what* systému. Po natrénování dokáže rozpoznat se 100 % úspěšností barvu i tvar obou objektů od velikosti 15x15 neuronů, ale chyba prostorového vztahu je 75%. jelikož informaci o poloze objektu nedostává. Tato chyba je následně přenášena do multimodální vrstvy. Možným řešením je použití asymetrických velikostí jednotlivých subsystému, kde je chyba eliminována pomocí větší velikosti auditivní unimodální vrstvy a vizuálního *where* systému. Provedli jsme v tomto směru několik úvodních pokusů, ale asymetrická verze obsahovala opět chybu v oblasti prostorových vztahů. Dalším cílem ve výzkumy je proto identifikace zdroje chyb jednotlivých subsystému a navržení architektury, která tyto nedostatky překonává. Technickou překážkou uvedených modelů je nedostatečná implementace algoritmů pro výpočet SOM map, které nejsou plně paralelizované. Při současné velikosti modelu, který obsahuje 6000 neuronů a 580000 propojení jsou výpočty na jednom jádře procesu velmi zdlouhavé a zaberou v průměru 150 hodin pro jednu vrstvu modelu. S dalším rozvojem modelu tedy bude nutné vytvořit paralelní způsob výpočtů neuronových vah.

5 Závěr

Prezentovaný model je příspěvkem do oblasti ukotvení symbolů. Jedná se o typ architektury založené na učení bez učitele, která má demonstrovat možnosti reprezentace znalostí založené pouze na vstupech z prostředí. Základní myšlenka vychází z předpokladu, že společný výskyt symbolického a subsymbolické vrstvy reprezentace v prostředí umožní jejich vzájemné mapování bez nutnosti apriorních znalostí. V rozšířené verzi by měl být systém schopný na základě učení odvodit základní formy gramatiky, schopnost mapovat abstraktní symbolické označení příslušnému konceptu v rámci hierarchických reprezentací a popřípadě vytvořeni pomocí procesů abstrakce a generalizace elementární formu logiky založenou na ukotvených reprezentacích.
V současné fázi bylo prokázáno, že je systém schopný mapovat jednotlivé slova věty na příslušné subsymbolické reprezentace bez apriorní znalosti. V dalších fázích vývoje by mělo dojít k vylepšení architektury v oblasti reprezentace prostorových vztahů a následné pokusy v oblasti reprezentace invariantních vlastností objektů. V rámci zvýšení plausibility modelu chceme rozdělit zpracování vizuálních vjemů na *what* systém, *where* systém a samostatné zpracování barev, doplnění o elementární pozornostní mechanismy a také reprezentační vrstvy založené na hierarchických sítích, pro zachycení abstraktních vlastností objektů, které mohou posloužit pro reprezentaci extra-geometrických vlastností objektů a jejich prostorových vztahů.

Poděkování

Tento projekt byl vyvíjen ve spolupráci s Doc. Igorem Farkašem (FMFI, KU Bratislava) a financován v rámci Národního stipendijního programu SAIA pod názvem „Multimodální reprezentace v oblasti vnímání prostoru“.

Literatura

[1] Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577-609.
[2] Coventry, K. (2004). *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions (Essays in Cognitive Psychology)*. Psychology Press.
[3] Dorffner G., Hentze M., Thurner G. (1996). A Connectionist Model of Categorization and Grounded Word Learning, *in Koster C., Wijnen F. (eds.): Proceedings of the Groningen Assembly on Language Acquisition (GALA '95)*.
[4] Farkas, I., Li, P. (2002). Modeling the development of lexicon with a growing self-organizing map. In H.J. Caulfield et al. (Eds.), Proceedings of the 6th Joint Conference on Information Sciences, Research Triangle Park, NC, pp. 553-556,.
[5] Harnad, S. (1990). The Symbol Grounding Problem. *Physica D*, 335-346.
[6] Herskovits, A. (1988). Spatial Expression and the Plasticity of Meaning. In: Rudzka-Ostyn, B. Topic in Cognitive Linguistics. John Benjamins Publishing Company, Amsterdam.
[7] Johnson-Laird, P.N., Miller, G.A. (1976). *Language and Perception*. Belknap Press.
[8] Kohonen, T.(1989). *Self-organization and associative memory*. New York: Springer.
[9] Landau, B., Hoffman, J.E. (2005): Parallels between spatial cognition and spatial language: Evidence from Williams syndrome. *Journal of Memory and Language,* 53(2):163-185.
[10] Paivio, A.(1986). *Mental representation: A dual coding approach*. New York: Oxford.
[11] Peirce, C.S.(1958). *Collected Papers of Charles Sanders Peirce*, vols. 1–6, Charles Hartshorne and Paul Weiss (eds.), vols. 7–8, Arthur W. Burks (ed.), Harvard University Press, Cambridge, MA, 1931–1935,.

[12] Saussure, F. de (1965). *Course In General Linguistics*. McGraw-Hill Humanities/Social Sciences/Languages.
[13] Tarski, A. (1944). The Semantic Conception of Truth: and the Foundations of Semantics'. *Philosophy and Phenomenological Research,* 4(3):341-376.
[14] Vavrečka, M. (2008). *(in Czech)* Application of the cognitive semantics in the model for the spatial terms representation. *Unpublished PhD thesis, Masaryk University in Brno, Czech Republic.*