The Language Mapping Phenomena

A computer simulation approach

Bc. Miroslav Vales

Cognitive Informatics Programme

University of Economics Prague

‣ MAIN TOPICS: Linguistic simulation, language evolution, multi-agent simulation, language mapping

‣ PREMISES: Human beings can handle different tasks using learned, established patterns that usually provide output good enough to one's standard. The patterns that are involved intuitively or explicitly. These acts of thinking and behavior are linked to the anticipated effects and may be perceived as the concepts having pragmatic, semiotic dimensions.

Human language is one of such tools. It has fundamental and unique position because it is necessary to keep formalization and standardization qualities among speakers to ensure its social, communication function. But the language is not limited to the interaction solely. It mutually influences human perception and it is possible to think of it backwards as well, whereas automatized language demonstrations are interpreted as an indicators of particular speaker's specifity.

This paper takes a multi-agent computer simulation as an instrument of research to reach limited insight in the field of language genesis and evolution. From this perspective, the world of language is modeled as a space of different speakers that use common words (strings) with defined probability in order to express agent's subjective meanings. Agents have a procreative power to generate new words and have the power of influence to codify new words, so any other speaker may communicate them then. They are also bound by the forgetting factor, which means the less used words can be forgotten.

‣ POINTS OF INTEREST: The motivation is to find behavior archetypes among mentioned factors and to describe general dynamics with regard to various set-ups. For instance simulation should provide answer on differences between small of speakers against the large one. Or a differences between the rigid language and the language environment of creative speakers.

‣ SIMULATION ABSTRACTION: The simulation uses the following abstraction - each speaker (an agent) has its own set of meanings located in two-dimensional space. Agent has a motivation to use some of these, but he can communicate them thru string labels only. So each agent has a meaning map and an overlapping labels (strings) map based on same principle. The interaction during communication between these two layers are affected by two main aspects - two-dimensional distance and labels similarity. It will be discussed lately.

Each agent has a set of attributes affecting his behavior. First of them is an activity* that indicates how many meanings agent tries to communicate during one round. Second is a creativity* which stands for the probability of creating a new label or editing an existing-one, if there is none available label. And third is agent impact*, which enables a possibility to sort the set of agents by their significance. Because from time to time there occurs two important events.

One of them is an update of an aggregate label map. This is a map created by averaging a label maps of the most significant agents (it is an analogy of a codification). And the second event is the harmonization, whereas the aggregated map is projected into the label map of particular agents (it is an analogy of language learning). Thus there is a suggestibility* (or learning factor) describing how easily an agent adopts labels from aggregate map. And finally, a forgetting factor* causes a continuous loss among personal labels that tops the number of inborn meanings.

* - a value of marked variable is limited by the interval [0, 100]

‣ DESCRIPTION IN DETAIL:

As has been mentioned before, each modeled speaker is an agent entity defined by a set of fixed or dynamic properties. These things predetermine the behavior in every simulation step.

It is known that human brains have the same structure in general, but in the scope of neurons and synapsis the structure is highly complex and makes any human being a unique individual. In case of simulation the agents have a fixed set of randomly generated meanings, where each meaning has its own 2D location and its value of importance. This first-level layer of valuable or less essential meanings establishes agent’s uniqueness as well as a need to communicate them. The question is how.

Agents use a dynamic set of labels (strings) to express their fixed meanings. Agent is influenced by his activity rate, a percent of meanings that should be expressed. Also each agent has defined creativity and forgetting rate to imitate such effects. The paradox is that the learned layer of labels is not necessary compatible with primary meaning layer. Expressing is done by comparing 2D positions of meaning and label. The nearest label is taken with regard to the position of source meaning. In case of worst distance, the creativity effect would be prosperous, ensuring better-balanced layers mapping.

But the personal use of language is just one of the sides. An agent have his own label (string) map that could be shared among other agents easily. Contrary the meaning layer is private and incommunicable part. From broader perspective, the agents can use their label maps based on initial, generated map and they are updating it step-by-step. The final outcome is aggregated, common label map, created by the merge of agents' maps according to the impact factor of each agent. And other way round the agents' maps are influenced by the global aggregated-one. This is the effect of language codification and language learning.

The most important agenda takes place in label manipulation. Agents must be specific with regard to label perception, thus each has the sensitivity factor (limited by the interval [0, 100]), a variable describing how agent's perception and label differentiation un/limited are.

It is necessary to select the main approaches, the language comparison tools to simulate and evaluate labels. In this case two methods were chosen - the Levenshtein distance is comparison metric expressing the difference amount of two strings. For instance, two strings "Prague" and "Plaguy" has two different characters {„l", "y"}, so the LD equals two (or 33.3 %). But the problem is such utilization does not reflect the phonetic aspect, there are a lot of differences between written and spoken. Thus appropriate way to handle this catch is to establish another filter transcribing label with respect to pronunciation. Double Metaphone algorithm does such work. It generates the pronunciation hash of word. Two differently written words can have the same pronunciation and hash. This solution makes possible to compare two or more labels using standard Levenshtein after applying Double Metaphone first.

The combination of Levenshtein and Metaphone methods is used to compare labels similarity, f. i. comparing a personal label with an aggregate label parameterized by the agent's sensitivity and the agent's weight of comparison methods.

‣ SIMULATION MECHANICS IN GENERAL: Described multi-agent model leads to the simulation, where following principles can be observed.

The system dynamic is constrained cardinally by the agent meanings set. Keeping this set fixed provides the main tendency to transform the personal label map in accordance with one's meanings. Contrary, the other fundamental impulse rises from the backwards influence of aggregate label map. Thus, these two rivalling pressures constitute the conflict that the language environment at an ideal state should handle.

The first phase of usual simulation is devoted to structural changes in agents' maps and in the aggregate map. Routinely the agents adopts new labels or generates better labels with regard to their meaning layer. It depends on the environment set-up, but the most relevant ones provide impulses that are transferred to the most suggestible ones via aggregate map.

The most important and distinguishable breakpoint of the simulation consists in getting environment in "the equilibrium state". It means the progression trend of notable parameters (number of labels, labels use, &tc.) has been stabilized despite

same level of volatility caused by the perpetual agents rivalry persists. Such situation indicates there was found more or less stable modus vivendi between the personal label maps of active agents and the aggregate, global one.

The labels of aggregate map is adopted by agents maps respecting the meanings allocation and suggestibility and sensitivity constrains. If there is a possibility to establish a new label its success on aggregated level is predetermined by the amount of agents (speakers) that recognize it as a following node of their own meaning layer. The procreative impulses take part from below to above. When the equilibrium state is reached, the rest of simulation is monotonous.

‣ RESULTS: Whole simulation abstraction highly evokes the comparison of the monopoly and the competition. In the case of strictly defined, rigid languages the speakers are bound by the global codification despites it may dictate them the less effective adaption of label structures connected to their inborn meaning layer. In the event of flexible language of creative speakers, the mapping can be more effective in the scope of particular agents, but the non-closed set of labels may bring out an unstable environment losing its main communication feature. (The 2D distance between source agent's meaning and the nearest label used for the communication can be utilized as the mentioned efficiency criterion.)

Re-simulating the many set-up cases has provided the following patterns. (1) Lower creativity of agents increases an average time needed to reach the equilibrium state and vice-versa. This is one of the most staring parameters taking its part between one's personal optimum and the optimum on global level, because the high creativity increases the final volatility of labels as well. (2) High suggestibility (learning) factor of agents decreases equilibrium-seeking time and vice-versa. But such tendency also influences an average lifetime of labels. In the case of personal map, the average label lifetime is lowered and in the event of aggregate map the effect is antagonistic. (3) Assigning the higher weight to the Double Metaphone against the Levenshtein in the frame of string similarity comparison results in lower equilibrium-seeking time and vice-versa. (4) The high amount of suggestible agents can absorb the indicators volatility. But this feature brings up a question how the two-dimensional space of meanings and labels should be calibrated. Because in performed simulations it was limited by the interval [0, 1000] and thus any increase in the number of agents leads to the creation of next overlapping maps.

‣ SUMMARY: This work predicates that multi-agent computer simulation can be a productive instrument in the field of linguistic research, providing another perspective. Double Metaphone and Levenshtein are essential methods enabling a possibility to imitate the text processing by human cognition.