Artificial Intelligence (A.I.) and Constructed-Language Emergence: overtaking Synthetic *Conlangs* with Self-Organized Linguistic Systems (SOLS)

Diego Gonzalez-Rodriguez
xmunch@xmunch.com, Group of Economics and Complexity, University of Valencia, Avinguda Tarongers, s/n, 46022, Valencia, España

Jose Rodolfo Hernandez-Carrion
rodolfo.hernandez@uv.es, Group of Economics and Complexity, University of Valencia, Avinguda Tarongers, s/n, 46022, Valencia, España

Abstract

This work is oriented to explore the potential of bottom-up generative processes in the context of *conlang* production, aiming to describe the basis of a new field of research: *Self-Organized Linguistic Systems* (*SOLS)*.

The *Self-Organized Linguistic Systems* approach provides a framework for the creation of self-generated artificial languages, emergent and dynamic alternatives to contemporary synthetic *conlangs* such as Esperanto, Toki Pona or Lojban, which have been precisely engineered with specific prefixed goals.

SOLS may serve as a starting point for the development of context-dependent or domain-specific languages which may be used for several purposes such as (a) second language acquisition, (b) communications between robots within unpredictable situations, (c) knowledge representation, (d) linguistic emergence in social agents, (e) visual language production or generative art.

Rather than relying on traditional Artificial Intelligence approaches to knowledge representation, this new field of research is specified under the perspective of both self-organized systems and constructed languages. It considers that the development of *conlangs* can happen in artificial societies of simple agents, that is, as the output of social interactions in computational simulations under the agent-based modelling paradigm. Emergent *conlangs* share some of the elements of natural languages and can provide a live and dynamic communication system for artificial agents in situations in which objects and actions are not pre-existent and in which lexicon should emerge from interactions within a shared environment.

This work starts contextualizing the notion of *language* and analyzing some of the already existing *constructed language*s. It explains historical and technical reasons for the design of *conlangs* and exposes a different theoretical approach.

This study also discusses how traditional ways of knowledge representation can be overtaken under the design of unconventional systems, and describes how language can be revisited with new systems of representation that go beyond the visual and audible options and become multidimensional.

Finally, this work introduces the notion of emergent languages and exposes the foundations of *Self-Organized Linguistic Systems. SOLS* rely on the theoretical framework of *Complex Adaptive Systems* in conjunction with the development of computational tools and simulations such as agent-based models.

SOLS can be helpful to understand better the nature of language and sense-making as emergent processes. This paper exposes possible implementations and further developments of the initially presented concepts. Rather than developing a linguistic theory or using existing frameworks such as Paul Hopper's functional work on emergent grammar, this research exposes some of the first steps in the development of straightforward algorithmic approaches for the generation of both descriptive and imperative lexicons in the context of agent-based models.

In the proposed initial *SOLS* model, automatic generation of lexicon takes place in the context of a digital environment with objects, actions and agents with embodied cognition through peer-to-peer interactions.

Specifically, this paper exposes how SOLS can be developed with bi-dimensional games and simulations. Non-interactive agent-based SOLS can allow artificial agents to independently evolve emergent languages as part of their self-organizing or adaptation processes. However, game-style SOLS can also provide an open environment in which artificial agents can also interact with human agents, leading to human-machine communications and peer learning processes.

In conclusion, *SOLS* are the intersection of artificial agent-based models with constructed languages, and may provide news ways for knowledge representation, artificial systems development and constructed languages design.

Keywords

artificial intelligence, self-organization, emergence, constructed languages, agent-based modelling, conlangs

About the Authors

Diego Gonzalez-Rodriguez is PhD in Information Science, MS in Artificial Intelligence and Engineer in Computer Systems. He has been Research Fellow at the P2P Lab, a spin-off of Tallin University of Technology and the P2P Foundation and has collaborated with research groups such as the Global Brain Institute of the Free University of Brussels, the Critical Making Lab at the University of Toronto, the University of North Carolina at Greensboro, the University Rey Juan Carlos and Technical University of Madrid in diverse topics such as Artificial Intelligence, Complex Adaptive Systems, Software as a Service, P2P social dynamics and the Social Implications of Technology. He has worked as Software Engineer for several organizations and companies and has been founder of diverse projects. Currently he also collaborates as a researcher with the Group of Economics and Complexity at the University of Valencia.

Jose Rodolfo Hernandez-Carrion is Associate Professor at the Department of Applied Economics at the University of Valencia UVEG (Spain), holding a Ph.D. in Economics. He has been teaching Economics 101 since 1989 in different centers and various languages (Catalan, Italian, Spanish and English) and publishing in international journals with an open and critical multidisciplinary approach. He served in the governing board of Spanish Association of Regional Science (AECR) and Spanish Society of Systems Science (SESGE), where he was Vice-President and Editor in Chief of RIS (Revista Internacional de Sistemas – International Journal of Systems). After his Ph.D. he has conducted research on “Economics Education for Change” and “New Pedagogies in Teaching Economics.” Connecting with Regional Economics, he was Visiting Scholar at University of Bologna (1992) and participated in the Icarus Project at the supercomputing center CINECA (Italy), later at University of Limerick (Ireland), University of California at Berkeley, UCLA (Los Angeles), University of North Carolina at Greensboro (UNCG), and UIUC in the United States (USA). He has been ERASMUS+ Professor invited four times to several European countries since 2013, being keynote speaker with his lecture: ’Economics today: dancing without borders’ during the International Week in Portugal and Poland.

1. Introduction

This work is oriented to explore the potential of *bottom-up* generative processes in the context of *constructed languages*, aiming to describe the basis of a new field of research: *Self-Organized Linguistic Systems* (*SOLS)*.

According to the Cambridge Dictionary [1], language is defined as "a system of communication consisting of sounds, words, and grammar" and "the system of communication used by people in a particular country or type of work". We consider these definitions both inaccurate and non-generalizable for artificial agents and artificial societies, in which the existence of sounds, words, grammar or human concepts such as "country" or "work" may not be necessarily present.

Rather, we conceive the notion of language as something broader which should be specified in in two different ways:

- (L1) Language is any set of rules which allow the representation of knowledge in a physical or virtual means, and its interpretation by sensible agents.
- (L2) Language can be used as a tool of communication between two or more sensible agents, allowing them to exchange knowledge according to a common protocol (which can be both pre-designed or emergent) and a set of perceptible elements.

Assuming both assertions as our starting point, we could define a simple taxonomy of languages in which we distinguish between *constructed* and *natural languages*.

Generally, we usually call *synthetic languages, artificial languages, constructed languages* or *conlangs* to those which have being designed or created with a pre-fixed purpose by a person or a group of people, as opposed to *natural languages,* those which have evolved historically in cultural, religious, tribal or national contexts [2,3,4].

According to that contraposition, we would classify Esperanto, Lojban, Ido, Toki Pona, Klingon or Newspeak as invented or *synthetic languages*, while we would conceive Spanish, English, Italian, Arabic, Hebrew, Japanese, Cherokee, Panjabi or German as *natural languages*.

However, this taxonomy is also inaccurate. Some of the languages that would be considered as *natural* have been also partially designed or historically constrained by individuals or institutions. We have the example of the Royal Academy of Spanish Language, which regulates and defines static rules in order to stablish the correct use of Spanish; or the *Euskaltzaindia,* which developed the Standard Basque language to unify several dialects spoken in Euskadi. We can also consider the scripts of *natural languages* as the result of artificial design. For example, the Korean alphabet or *Hangul,* was devised by King Sejong in 1443; another example is the Serbo-Croatian script or Gaj's Latin alphabet, which was created by Ljudevit Gaj.

In the same way, languages which are conceived as invented, artificial or *synthetic,* can also present characteristics of natural languages. For example, Esperanto has developed their own culture and has dynamically evolved since 1887, going beyond its original rules and vocabulary. Currently, Esperanto has a large community of native speakers. According to

Ethnologue [5] there are from hundreds to thousands of *denaskuloj*, Esperanto native speakers. Therefore, we can say that artificial languages, eventually, may end up evolving and developing characteristics of natural languages.

According to this, the conceptual boundary which separates both realms is quite permeable, considering that:

- Sensitive agents in both artificial and actual societies evolve dynamically as constitutive parts of Complex Adaptive Systems, as it has been explored in our previous research [6, 7].
- Any language used by a group of social agents will eventually have an evolutionary component unless this is constrained by design. And in that case the constraint can lead to a fork of the linguistic community (see Ido, a fork of Esperanto).

That permeability leads us to oppose to any qualitative taxonomy which pretends the domination of natural languages over artificial languages or the opposite. We consider that languages, independently of any taxonomy, should serve the purpose described previously (see L1): allow the representation of knowledge in a physical or virtual means and its interpretation by sensible agents. Additionally, languages can be used as tools for communication between two or more sensible agents, allowing the exchange of knowledge according to a pre-designed or emergent common protocol and a set of perceptible elements (L2).

The nature of sensible agents it is not specified in neither L1 or L2, so they can be human beings, animals, electronic devices or software entities. The only requirement for them is to be *sensitive:* having senses, sensors or functions to identify and distinguish perceptible elements which can represent different forms of data.

We could design a software platform in which different agents with artificial intelligence interact autonomously. Those agents themselves may create their own set of perceptible elements and rules for communication through P2P interactions. And that emergent set, according to the above-mentioned concepts, would be considered a valid language if it serves for the purposes of language (L1 and L2).

Constructed or *synthetic languages* have been designed with diverse and specific purposes. For example, Lojban provides high levels of accuracy allowing a *first order predicate logic* communication [8]. Ben Goertzel proposed a improved version of this language, Lojban++, as an interlingua for communication between humans and an Artificial General Intelligence [9]. Some artificial languages can be specifically useful for Embodied Natural Language Processing [10] and some of them are used for their propaedeutic value and can serve as tools to understand the cognitive mechanisms of second language acquisition [11]. Some other *conlangs* have been designed with mere artistic reasons [12].

This wide and diverse ecosystem, in which different languages are developed according to historical or technical reasons, presenting their own advantages and disadvantages -and in which the synthetic nature of *conlangs* can not be isolated of a natural evolution through time-, leads us to a scenario in which computer programs themselves can create their own languages based on specific design constraints or purposes, and with unconventional and multidimensional properties.

In that sense, the *Self-Organized Linguistic Systems* approach provides a theoretical framework for the creation of self-generated languages in computational environments, dynamic sets of rules and perceptible elements for both knowledge representation (L1) and communication purposes (L2): *emergent conlangs*.

Self-Organized Linguistic Systems (*SOLS)* assume the evolutionary component of linguistic production, and can be used as a theoretical framework for the development of context-dependent or domain-specific languages for several purposes such as (a) second language acquisition, (b) communications between robots within unpredictable situations, (c) knowledge representation, (d) linguistic emergence in social agents, (e) visual language production or (f) generative art.

2. Self-organized linguistic systems

Rather than relying on traditional Artificial Intelligence approaches to knowledge representation, *Self-Organized Linguistic Systems (SOLS)* are specified under the perspective of both self-organized systems and constructed languages. They can also include a component which allows the visualization of the generative process.

This approach considers that the development of *conlangs* can happen in artificial societies of simple agents, that is, as the output of social interactions in computational simulations under the agent-based modelling paradigm. *Emergent conlangs* share some of the elements of natural languages and can provide a live and dynamic communication system for artificial agents in situations in which objects and actions are not pre-existent and in which lexicon should emerge from interactions within a shared environment.

SOLS rely on the theoretical framework of *Complex Adaptive Systems* in conjunction with the development of computational tools and simulations such as agent-based models.

SOLS can be helpful to understand better the nature of language and sense-making as emergent processes. This paper exposes initial implementations of the tools which will be used for this purpose, as an introduction to further developments of the initially presented concepts. Rather than developing a linguistic theory or using existing frameworks such as Paul Hopper’s functional work on emergent grammar, this research is more general and exposes some of the first steps in the development of straightforward algorithmic approaches for the generation of both descriptive and imperative lexicons in the context of agent-based models.

In the proposed initial *SOLS* model, automatic generation of lexicon takes place in the context of a digital environment with objects, actions and agents with embodied cognition through peer-to-peer interactions. So in order to develop SOLS, we initially need:

- An internal knowledge representation framework which will operate internally, as part of the cognitive processes of each agent. It should provide the visualization of the knowledge base evolution in real time, preferably in 3D to browse better the network of triplets and relationships between nodes.

- A simulation environment for P2P social dynamics: this multi-agent simulation tool should provide a visual component and the agents should communicate to each other through perceptible elements. The simulator can be bi-dimensional.

Specifically, this paper exposes those initial tools which can lead to the development of proof-of-concept SOLS based on bi-dimensional games and simulations with a 3D knowledge representation visualization.

Non-interactive agent-based SOLS can allow artificial agents to independently evolve emergent languages as part of their self-organizing or adaptation processes. However, game-style SOLS can also provide an open environment in which artificial agents can also interact with human agents, leading to human-machine communications and peer learning processes.

3. Knowledge representation

Traditional ways of knowledge representation can be overtaken under the design of unconventional systems, and language can be revisited with new systems of representation that go beyond the visual and audible options and become multidimensional.

Sign languages, music-based languages like *Solresol*, or color-based alphabets like *Colorbet* has been deeply explored during the last centuries as visual and audible alternatives to traditional languages. Some artistic explorations have been also made under the label of *constructed scripts or con-script* (see Fig.1).

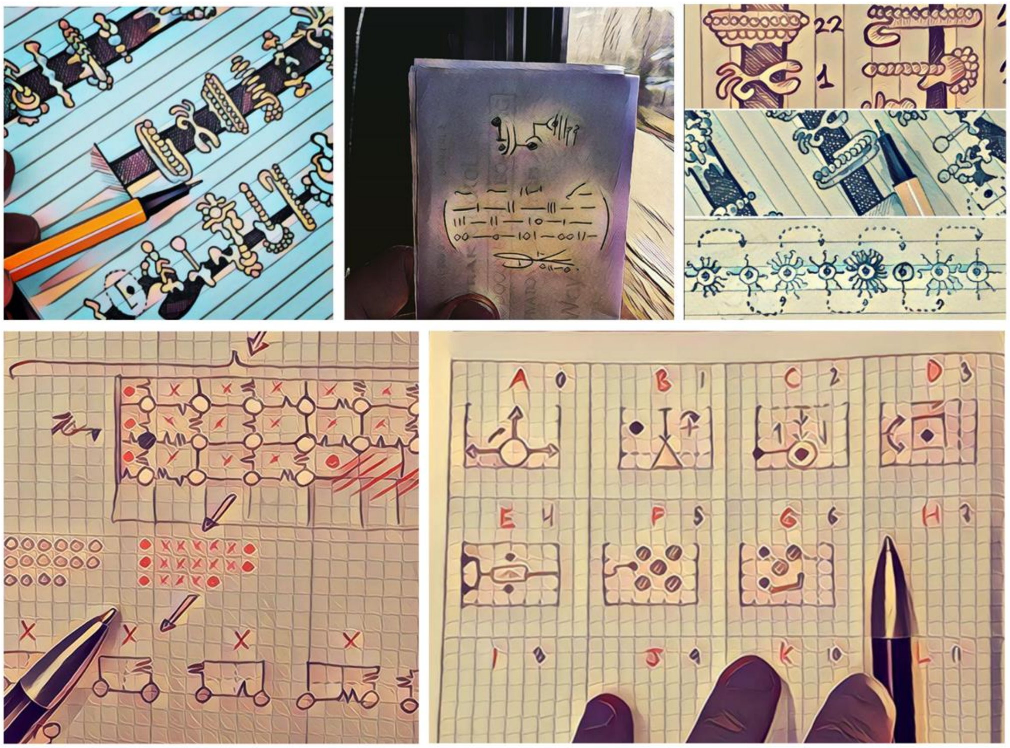

Figure 1. *Some examples of visual representation systems for artificial languages based on shapes and interaction rules.*

Our own tool for knowledge representation in artificial agents, it is the *xmunch-atomspace* or *XA* [13], which allows the dynamic construction of neural and semantic networks for knowledge representation, reasoning and learning in an abstract space of atoms (vertices and edges). It intends to be a tool for helping in narrow Artificial Intelligence, but also to be used in the implementation of cognitive architectures for the Artificial General Intelligence (AGI) community and used by artificial agents in language production tasks.

XA implements an interface to communicate in real time with an Ubigraph server. This feature allows researchers and developers to visualize both declarative and procedural knowledge. With ***XA*** it is very easy to visualize both sub-symbolic representations or symbolic/semantic relations in real time, but also learning algorithms, atom-based state-machines and spreading activation. In Fig. 2, we display some of the relational visualizations of *XA*.

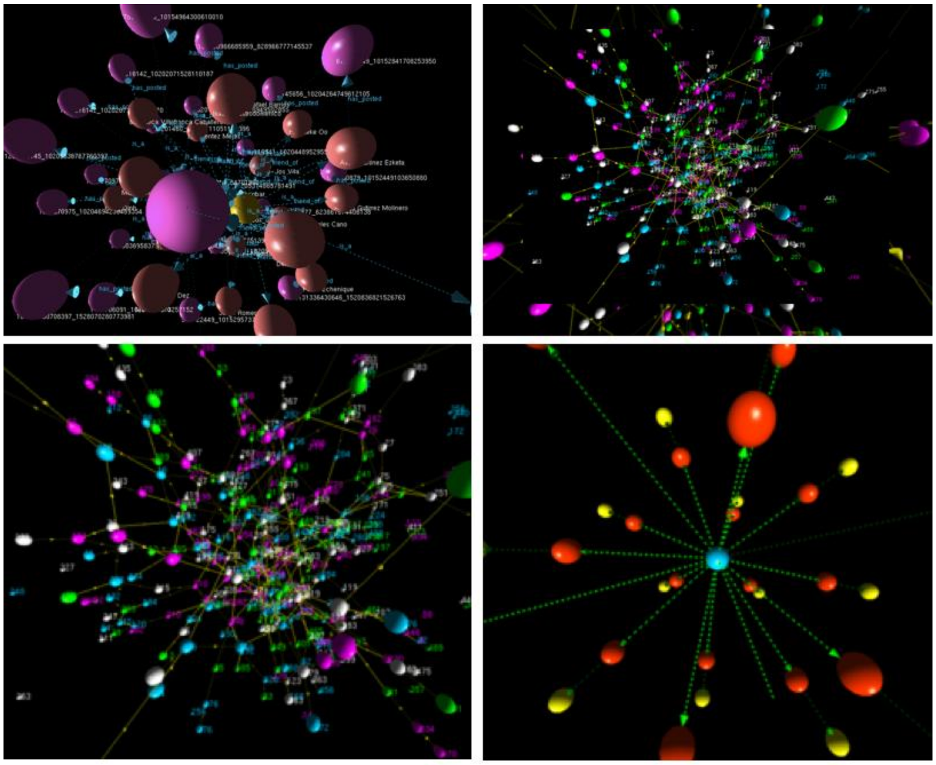

Figure 2. *Visual representation of knowledge using Ubigraph and xmunch-atomspace.*

The persistence layer of *XA* is written *in xa-language*, an interoperable human and machine understandable language. It is a simple language with a small set of rules which can be used to create new atoms in any running instance of **XA** by using the *AtomSpace REST Parser* or similar tools.

The basic rules of *xa-languages* are:

1. Every predicate in *xa-language* is written in a new line.

2. Every predicate in *xa-language* follows an A or B *xa-language predicate* structure.

3. Every *A xa-language predicate* follows this structure:

   NAME is_a TYPE [PARAM_1:VALUE_1 ... PARAM_N:VALUE_N]

   - Examples:

   - Susan is_a person student:yes age:23 gender:woman
   - Peter is_a person age:25 gender:man children:3
   - Carolina is_a person
   - Madrid is_a city geo-x:40.383333 geo-y:-3.716667

- Europe is_a continent
  Barcelona is_a city has_beach:yes

4. Every *B xa-language predicate* follows this structure:

   NODE_NAME_1 RELATION NODE_NAME_2 [PARAM_1:VALUE_1 ... PARAM_N:VALUE_N]

   - Examples:

   - Peter is_friend_of Carolina affinity:0.5
   - Madrid is_located_in Europe
     Madrid is_connected_by_road_to Barcelona distance:600 highways:yes

5. Every *B xa-language predicate* requires that both **NODE_NAME_1** and **NODE_NAME_2**have been declared with two *A xa-language predicates* before.

6. Elements **NAME, TYPE, NODE_NAME, RELATION, PARAM** or **VALUE** can not contain empty spaces ( ) or double quotes ("). Spaces should be replaced by underscore lines (_). Single quotes (') are allowed.

XA and the set of tools developed around it are an internal knowledge representation framework for our cognitive agents. Its REST API allows the visualization of the knowledge base evolution in real time, providing a better understanding of the network of triplets and relationships between nodes.

4. Multi-agent simulation

Additionally to our knowledge representation tool, we also need a simulation environment for P2P social dynamics. As it has been said, this multi-agent simulation tool should also provide a visual component and the agents should communicate to each other through perceptible elements. We have developed the *SciArt 2D Simulator* [14] as a 2D Graphics Engine developed to be used as an extensible tool for our research in agent-based modeling, artificial life and educational games.

Our simulator is developed in *Groovy, Java* and *Processing* and it provides basic tools to model multi-agent systems, social behavior and emergent properties in digital environments. It will serve as a support to our study of *Self-Organized Linguistic Systems (SOLS)* in a bi-dimensional environment and will provide an easy way to develop from simple Turtle-like systems for educational purposes to complex *P2P Social Dynamics* (see Fig. 3).

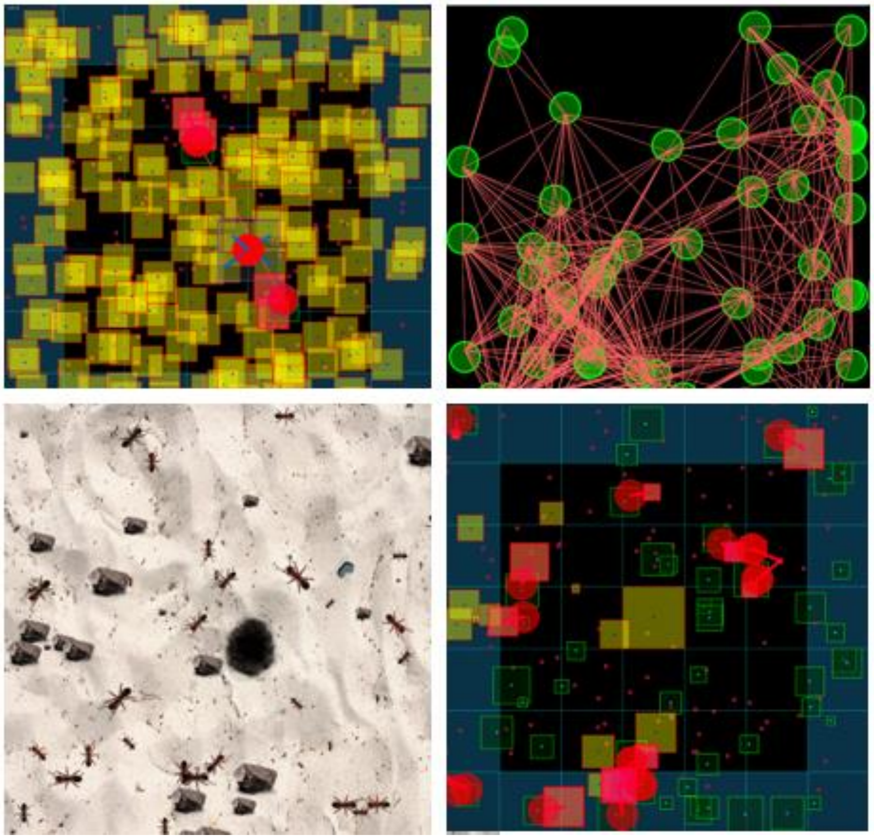

Figure 3. *Multi-agent models implemented in the SciArt Simulator*

5. Conclusion

The *Self-Organized Linguistic Systems* (SOLS) are a theoretical framework for the creation of self-generated artificial languages, emergent and dynamic alternatives to contemporary synthetic *conlangs* such as Esperanto, Toki Pona or Lojban.

SOLS will be the intersection of artificial agent-based models with constructed languages, providing news ways for artificial systems development and constructed languages design.

For the development of actual implementations of this theoretical framework, we have developed both modeling and visualization tools according to the L1 and the L2 components of our definition of language:

- (L1) Language is any set of rules which allow the representation of knowledge in a physical or virtual means, and its interpretation by sensible agents.
- (L2) Language can be used as a tool of communication between two or more sensible agents, allowing them to exchange knowledge according to a common protocol (which can be both pre-designed or emergent) and a set of perceptible elements.

An initial work has been done with the *xmunch-atomspace* and the *SciArt simulator*, which constitute the initial implementations of both our knowledge representation toolbox and our bi-dimensional simulator of P2P Social Dynamics.

This work is just a starting point which eventually will allow us to model agent communications in a language production environment. But we also would like to present SOLS as an inspiration for the *conlang* community, as a theoretical trigger to stimulate the development of computational tools for future *emergent conlangs.*

References

[1] Language Meaning in the Cambridge English Dictionary. (n.d.). Retrieved November 19, 2016, from http://dictionary.cambridge.org/dictionary/english/language

[2] Kuhn, T. (2016). The Controlled Natural Language of Randall Munroe’s Thing Explainer. In B. Davis, G. J. Pace, &amp; A. Wyner (Eds.), Controlled Natural Language (pp. 102–110). Springer International Publishing. https://doi.org/10.1007/978-3- 319-41498- 0_10

[3] Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. Computational Linguistics, 40(1), 121–170. https://doi.org/10.1162/COLI_a_00168

[4] Gobbo, F., &amp; Durnová, H. (2014). From universal to programming languages. Informal Proceedings of Computability in Europe.

[5] Esperanto. (n.d.). Retrieved November 19, 2016, from https://www.ethnologue.com/language/epo

[6] Gonzalez-Rodriguez, D., &amp; Kostakis, V. (2015). Information literacy and peer-to-peer infrastructures: An autopoietic perspective. Telematics and Informatics, 32, 586–593. https://doi.org/10.1016/j.tele.2015.01.001

[7] Gonzalez-Rodriguez, D., &amp; Hernandez-Carrion, J. R. (2014). A Bacterial-Based Algorithm to Simulate Complex Adaptive Systems. Lecture Notes in Artificial Intelligence (LNAI), 8575, 250–259. https://doi.org/10.1007/978-3- 319-08864- 8_24

[8] Speer, R., &amp; Havasi, C. (2004). Meeting the Computer Halfway: Language Processing in the Artificial Language Lojban. In Proceedings of MIT Student Oxygen Conference.

[9] Goertzel, B. (2013). Lojban++: an interlingua for communication between humans and AGIs. Artificial General Intelligence, 7999. https://doi.org/10.1007/978-3- 642-39521- 5

[10] Goertzel, B., Pennachin, C., Araujo, S., Silva, F., Queiroz, M., Lian, R., … Senna, A.(2010). A General Intelligence Oriented Architecture for Embodied Natural Language

Processing. Proceedings of the 3d Conference on Artificial General Intelligence (AGI-10),1–6. https://doi.org/10.2991/agi.2010.16

[11] Tily, H., Frank, M. C., &amp; Jaeger, T. F. (2011). The learnability of constructed languages reflects typological patterns. In Proceedings of the 33rd annual conference of the cognitive science society (pp. 1364–1369). Citeseer.

[12] Isenberg, R. (2000). Universitetet i Bergen. Artificial Languages. Retrieved September 2016, from http://folk.uib.no/hnohf/artlang.htm

[13] SciArt Lab. (2016). xmunch-atomspace. Hacking Science, Art and Technology. Retrieved October 2016, from http://www.sciartlab.com/?page_id=505

[14] SciArt Lab. (2016). 2D Simulator. Hacking Science, Art and Technology. Retrieved October 2016, from http://www.sciartlab.com/?page_id=517