Ontology Model of Language Evolution

Natalia V. MAKUSHEVA, Irina L. ARTEMIEVA, and Viktoriya L. ZAVYALOVA

1School of Natural Sciences, Far Eastern Federal University, Vladivostok, Russian Federation
2School of Regional and International Studies, Far Eastern Federal University, Vladivostok, Russian Federation

Keywords: Language evolution, Comparative linguistics, Evolutionary modeling, Software development.

Abstract. Our purpose is to design a mathematical model of language evolution among the language using community and develop the software system modeling the evolution process. The software main task is to design individual historical scenarios of particular languages, language groups and families, and show the quantitative and qualitative changes in the grammatical and lexical structure of language over time, depending on the social organization of a group of language learners and users.

Introduction

The methods of establishing the linguistic affinity use some empirical estimates of the rate of language change, based on comparison of lexical lists. But sometimes this kind of estimates does not provide the required accuracy, which significantly complicates the work of experts. This is because glottochronology can rely only on processes with a more or less constant rate. But in different socio-cultural environments, the same processes can take place in different ways. And the conclusions, which are valid for one epoch or culture, could be incorrect for another.

The purpose of our work is to develop a software product modeling language evolution that can give a significant refinement for evaluating the rate of change of language by creating individual historical scenarios for the development of individual languages, language groups and families.

To develop a model of language evolution, we chose an ontological approach that would allow us to design the modular structure [2, 6]. This, in turn, will allow the model to be used for verification of various hypotheses of the evolution of a natural language and quickly adapt it to new knowledge and data that appear in this area every day.

Language Evolution Modeling System

The natural language of the species is defined as a way of structuring and transmitting information within a group of individuals of this species.

The idea of the model is that the language can be represented in the form of a resource, consisting of a set of structural elements and distributed in the population of speakers. The structural elements of the language are embodied in the computer model by data structures composed in a special way.

To implement the basic model, we chose the following structural elements of the language:

• Words - formal-semantic units that have their own form and own meaning, organized in the way that when the form is violated, the value is lost, and the constituent part of the concept embodied in the word cannot be expressed by a part of the word;
• Types of words - labels of words, characterizing the possibilities and restrictions of the use of words in the composition of syntactic sequences;
• Sequences of types of words - syntactic structures, any kind of sequences.

In this paper we present fragments of developed model, which describes the individual's language structure, the message structure, and the evolution of the individual's language. This fragment defines
the representation of the language of the individual only. A formal statement is accompanied by a comment for understanding. The model presented in the multi-sorted language of logic [6].

Sort word: \{ \} N \setminus \emptyset An unordered set of words.
Sort type of word: \{ \} N \setminus \emptyset An unordered set of word types.
Sort types of word: word \rightarrow \{ \} type of word Each word corresponds to a finite set of types.
Sort words of the type: type of word \rightarrow \{ \} word Each type of words corresponds to a finite set of words.

sentence \rightarrow seq word The set of all possible (and impossible) sequences of words.
word type sequence \rightarrow seq type of word The set of all possible (and impossible) sequences of types of words.

Sort valid type sequences: \{ \} word type sequence A subset of all admissible sequences of types of words.

language element \rightarrow \{ word, type of word, word type sequence\} Structural elements of the language.

Sort individual’s words: moment X individual \rightarrow \{ \} word All the words of the individual at each moment of time.

Sort individual’s types of words: moment X individual \rightarrow \{ \} type of word All types of words of the individual at each moment of time.

Sort individual’s word type sequence: moment X individual \rightarrow \{ \} word type sequence All sequences of types of words of the individual at each moment of time.

Although studies in the field of brain mapping are only underway, there is reason to suppose that this language structure has a registered neurobiological basis [3, 5]. At this stage of the model development, we do not consider such a property of the language as inflectionality, but attention will certainly be paid to it in further research. The life of speakers (in the model) proceeds in discrete time. Speakers (individuals) can acquire a language - accumulate a language resource, and communicate with each other - transfer the language resource to each other by copying. According to recent research, adults acquire language using the same unconscious mechanisms that are involved in the primary language acquisition [4]. With the presentation of a new word (or pseudo-word), a memory trace (a group of neurons response to the presentation of an appropriate stimulus) is formed in the brain. Each new access to the memory trace activates it and increases the efficiency of the response. In this case, the activation of the memory trace (the structural node of the brain) associated with the word occurs not only with the active use (pronouncing the word), but also with the passive presentation of the word, even when the listener does not.

In order to model this process, in the individual language of the speaker (the recognizable set of elements) we define the passive (understood) and active (used in speech) subset, so that the active subset is included into the passive subset, and the passive subset is in turn included into the set of individual language. The inclusion of a structural element of a language into a particular subset is characterized by the value of the counter of its presentation.

Sort individual’s language: individual X moment \rightarrow \{ \} (\{ word XI [1, \infty) XI [1, \infty) XI [1, \infty) \}, \{\} \{ word, word \} XI [1, \infty) \}, \{\} \{ type of word XI [1, \infty) XI [1, \infty) XI [1, \infty) \}, \{\} \{ type of word, type of word \} XI [1, \infty) \}, \{\} \{ word type sequence XI [1, \infty) XI [1, \infty) XI [1, \infty) \}, \{\} \{ word type sequence, word type sequence \} XI [1, \infty) \}) The totality of all words, types of words, sequences of word types, counters of repetitions of language elements and their pairs, and inclusion thresholds for each individual at each moment of time.

Sort individual’s language next moment: individual X moment \rightarrow \{ \} (\{ word XI [1, \infty) XI [1, \infty) XI [1, \infty) \}, \{\} \{ word, word \} XI [1, \infty) \}, \{\} \{ type of word XI [1, \infty) XI [1, \infty) XI [1, \infty) \}, \{\} \{ type of word, type of word \} XI [1, \infty) \}, \{\} \{ word type sequence XI [1, \infty) XI [1, \infty) XI [1, \infty) \}, \{\} \{ word type sequence, word type sequence \} XI [1, \infty) \})
Sort threshold for inclusion into the passive language: individual \( X \) moment \( X \) language element \( \rightarrow I \{1, \infty\} \) The number characterizing the inclusion of a language element in a subset of the language of the individual that is used to generate messages.

Sort individual’s passive language: individual \( X \) moment \( \rightarrow \) \( \{\}{(\text{word} X I[1, \infty] X I[1, \infty] X I[1, \infty]), \{\}{(\text{word, word}) X I[1, \infty]), \{\}{(\text{type of word} X I[1, \infty] X I[1, \infty] X I[1, \infty]), \{\}{(\text{type of word, type sequence}) X I[1, \infty])})\} (v: \text{language element}) \) counter of occurrence of language element \( v \) \( \rightarrow \) threshold for inclusion into the passive language \( v \) A subset of an individual's language at each moment in time, in which all relevant elements of the language can participate in understanding messages; a totality of all words, types of words, sequences of types of words, repeat counters of which exceed the thresholds of inclusion in the passive language for each individual at each moment of time.

Sort threshold for inclusion into the active language: individual \( X \) moment \( X \) language element \( \rightarrow I \{1, \infty\} \) The number characterizing the inclusion of a language element in a subset of the language of the individual used to generate messages.

Sort individual’s active language: individual \( X \) moment \( \rightarrow \) \( \{\}{(\text{word} X I[1, \infty] X I[1, \infty] X I[1, \infty]), \{\}{(\text{word, word}) X I[1, \infty]), \{\}{(\text{type of word} X I[1, \infty] X I[1, \infty] X I[1, \infty]), \{\}{(\text{type of word, type sequence}) X I[1, \infty])})\} (v: \text{language element}) \) counter of occurrence of language element \( v \) \( \rightarrow \) threshold for inclusion into the active language \( v \) A subset of an individual’s language at each moment in time in which all relevant elements of the language can participate in messages generation; a totality of all words, types of words, sequences of types of words, repeat counters of which exceed the thresholds of inclusion in the active language for each individual at each moment of time.

Individuals in the model are not able to forget the language, i.e. they cannot lose resources, except when the individual “dies” and his/her resource is completely lost. In this model, we have not yet considered cases of severe brain damage (including age), when access to the language is lost partially or completely.

According to psycholinguistic theories, message processing consists of four steps [7]. The first step is the processing of phonetic, graphical or other information that encodes the utterance. The next step is lexical processing, the division of the utterance into lexical units. The third step is the semantic processing: at this step the lexical units distinguished in the utterance are associated with certain elementary concepts. The fourth step is the grammatical analysis of the utterance: at this step the connections between lexical units are established.

In our model, we retain these steps with some amendments. Because elements of the language are implemented by data structures, the individual speaker in the model does not need to process incoming information and recognize lexical units. In the basic model, we accept the hypothesis that all individuals have sufficient recognition ability. Violation of the recognition ability will be considered in further studies.

\[
((v1: \text{individual} v2: \text{moment} v5: \text{processed message} (v1, v2) v6: \text{repeats of language elements} (v1, v2) v3 \in I [1, \text{length (processed message} (v1, v2))]) v4: \text{individual’s language} (v1, v2) \text{ message receiving} (v1, v2) = \text{message receiving true, message understanding} = \text{message understanding true}) \Rightarrow \text{individual’s language next moment} (v1, v2) = \{\}{(\text{word} X (\pi(2, \pi(1, v4)) + \pi(2, \pi(1, v6)))) X I [1, \infty])}\} (v: \text{language element}) \]
receiving \((v1, v2) = \) message receiving true, message understanding \(= \{ \) message understanding false, message understanding partial \(\}) \Rightarrow \) individual’s language next moment \((v1, v2) = \{ \}\) ({}(word \(\cup \pi (1, \pi (1, v6)) X (\pi (2, \pi (1, v4)) + \pi (2, \pi (1, v6))) X I [1, \infty) X I [1, \infty) , \}\) (\{word, word\} X (\pi (2, \pi (2, v4)) + \pi (2, \pi (2, v6))))), \}\) (\{types of word \(\cup \pi (1, \pi (3, v6)) X (\pi (2, \pi (3, v4)) + \pi (2, \pi (3, v6))) X I [1, \infty) X I [1, \infty) , \}\) (\{types of word, types of word\} X (\pi (2, \pi (4, v4)) + \pi (2, \pi (4, v6))))), \}\) (\{word type sequence \(\cup \pi (1, \pi (5, v6)) X (\pi (2, \pi (5, v4)) + \pi (2, \pi (5, v6))) X I [1, \infty) X I [1, \infty) , \}\) (\{word type sequence, word type sequence\} X (\pi (2, \pi (6, v4)) + \pi (2, \pi (6, v6))))) Change the language of the individual after receiving the message.

Communication, i.e. the transfer of a resource from an individual to an individual is provided selectively.

Messages of individuals are constructed randomly, taking into account the frequency and connections of the structural units of the individual language [1], and also taking into account the modeled labor activity, social status and circle of communication (groups of individuals who can receive a message from this individual).

In the process of communication, errors and failures may occur. All this leads to a change in the qualitative and quantitative distribution of the language resource in the population.

The population can reach several tens of thousands of individuals, thus, it is impossible to manually enter the properties of each individual. To form a population of individuals with a common language, the values of the properties of the entire population and the properties of the population language can be entered directly through the interface or uploaded from the file. The software system, using this data, generates parameter values for each individual for the start-moment of the simulation, and also generates a simulation algorithm.

A set of characteristics of the population and language constitute a modeling profile.

Modeling the evolutionary process, the software system generates data in accordance with the modeling profile, i.e. forms a set (population) of individuals, each of which has its own language resource, and then, by step-by-step passage, calculates the state of the model population of speakers according to the parameters specified in the modeling profile.

Step-by-step passage. Modeling is provided by changing the state of the model population according to the parameters of the modeling profile.

Each step of the model includes a number of actions of the software system.

1. Removal of individuals with the status of life “dying”.
2. Selecting individuals (random) to assign the status of “dying” in the next step.
3. Creation of new individuals.
4. Generation of social statuses of new individuals according to data on the social structure of the population (if the social structure has not changed). Generation of new social statuses for all individuals according to data on the social structure of the population (if the social structure has changed).
5. The inclusion of new individuals into the circles of communication of already existing individuals, the creation of circles of communication of new individuals, the recalculation of communication circles for some already existing individuals (if the social structure has not changed). Recalculation of communication circles for all individuals taking into account the new social structure (if the social structure has changed).
6. Assignment of types of labor activity to individuals who have reached “working age”.
7. Recalculation of the circles of communication of individuals to whom the type of labor activity was assigned in account of this type of activity.
8. Change the communication status of individuals in accordance with the specified frequency of messaging.

Communication. Step Communication includes the following actions:

For those individuals whose communication status has changed to send a message, the following actions are involved:
1. Generation of the message using the active language of the sender.
2. Distortion of the message at the time of sending (may be zero).
3. Transmission of the message by the individual-sender to the individual-receiver.

For those individuals whose communication status has changed to receive a message, the following actions are involved:

1. Reception of the message by the individual-receiver.
2. Distortion of the message at the time of receiving (may be zero).
3. Processing of the message using the passive language of the individual-receiver.
4. Change of the language of the individual-receiver according to the language data received in the message.

Steps of algorithms 2 make it possible to create a noisy communication. The occurrence of errors can be either stochastic or structured.

Conclusions

The paper describes the project of a software system for creating individual scenarios for the occurrence of changes in languages, language groups and families. This will give researchers in the field of comparative linguistics an additional tool for studying language processes. In addition, such a system can be used by experts in the field of history, sociology, as well as in the educational process.

Currently, a prototype of the software system has been developed, with the help of which experiments have been performed to confirm the model's performance. The work is conducted jointly with linguists who act as experts and can assess the correctness of the model and the correctness of the

Further research will focus on refinement of the model, the introduction of new additional concepts, the formulation of relationships and connections of concepts with each other. It is also necessary to adapt the model to the requirements of researchers in the field of comparative linguistics, which implies the development of additional ontology modules describing different systems of knowledge and representations. In the future, this will create the opportunity to conduct a comparative analysis of linguistic hypotheses through a model experiment.

References


