Basic Assumptions about a Sign's Life Cycle for Mathematical Modelling of Language System Evolution

A.A.Polikarpov D.V.Khmelev

24 August 2000

KEYWORDS: language evolution, mathematical modelling, branching processes

AFFILIATION: Lomonosov Moscow State University, Russia; Heriot-Watt University, Edinburgh, UK and Isaac Newton Institute for Mathematical Sciences, Cambridge, UK.

POSTAL ADDRESS: 20 Clarkson Road, Cambridge, CB3 0EH, U.K.

FAX NUMBER: (01223) 330508

E-MAIL ADDRESS: polikarp@philol.msu.ru, D.Khmelev@newton.cam.ac.uk

Contents

1 Aim of the paper.

Modelling of language evolution should be based on some assumptions concerning its micro-level, i.e. level of its micro-units' development. A sign (morphemic, lexemic and phraseologic) is an elementary (micro-) unit on some certain level of language organisation. A sign's polysemy evolution in time is the most fundamental ontological fact. That is why it has become the starting point in the building of the mathematical model for the development in time and synchronic correlation of the whole system of a sign's features. Correspondingly, it can lead to building a theory of the organisation and historical development of language systems as a whole.

2 Basic assumptions.

(1) A sign's polysemy development is a branching process of generating new meanings from previously acquired (and, correspondingly, losing some previously generated) ones.

(2) According to the increase of the ordinal number i of meaning's generation within a sign there should proportionally grow the average degree of meaning's abstractness A_i (or, in other words, decrease the average degree of meaning's filling by some number of semantic components B_i). This means that $A_i = 1/B_i$.

(3) The more abstract, on the average, the meanings of some generation of a sign are, the greater stability L_i (length of life) specific to each of them.

(4) The more abstract, on the average, each meaning of some generation of a sign, the lower generating activity G_i (number of meanings of the next generation produced from a meaning in its life) specific to each of them is.

(5) The more abstract meanings of some generation, the greater sense volume V_i (number of senses covered by each of them) that is specific, on the average, to each of them.

(6) The greater sense volume of meanings of some generation, the higher, on the average, the frequency of use U_i for each of them is.

These assumptions provide us with the ability to draw some useful conclusions for modelling of some other functional dependences for any language sign, as well as for ensembles of them, i.e. for a language system as a whole.

3 3. Dying out, giving birth, and preserving of sign's meanings.

Consequence 1. From the fact of a finite number of features in any sign's meaning it follows that maximal possible number of generations of meanings in a sign can not exceed some n.
Consequence 2. From assumptions (1)-(4) it follows that $L_1 \leq L_2 \leq \cdots \leq L_n$ and $G_1 \geq G_2 \geq \dots G_{n-1}$.

We shall consider evolution of a sign in continuous time. Let $\gamma_i = 1/L_i$ and $\beta_i = G_i/L_i$. Clearly, γ_i is a decay rate of meanings of generation i in a sign and β_i is a rate for generating new meanings (meanings of the next generation $i+1$) by each meaning of a generation i. Let us assume that during small intervals of time Δt every meaning of a generation i independently of all other sign meanings does the following:

1) dies with probability $\gamma_i \Delta t + o(\Delta t)$,

2) if $1 \leq i \leq n-1$ then it generates a meaning of the next generation $i+1$ with probability $\beta_i \Delta t + o(\Delta t)$.

Otherwise a meaning just preserves itself, continues its existence (with probability $1 - (\gamma_i + \beta_i)\Delta t + o(\Delta t)$).

It is easy to prove that within the model activity of a meaning belonging to a generation i, i.e. average number of meanings of a generation $i+1$ produced by a meaning of a generation i, equals G_i.

4 Drawing a curve for a sign's polysemy dynamics.

It is much more difficult to check a conclusion that a polysemy curve of a typical sign should have only one global maximum. Assumption (7) reduces the model to the branching process with the definite number of types of particles (see [1]). Denote [1] by P^i_σ the probability of the fact that a meaning of a generation i will generate for the time t meanings determined by the vector $\sigma = (\sigma_1, \ldots, \sigma_n)$: σ_1 meanings of a generation 1, …, σ_n meanings of a generation n. Define generating functions

$$F^i(s) = \sum_{\sigma \geq 0} P^i_\sigma(t) s^\sigma,$$

where $s^\sigma = s_1^{\sigma_1} \times \cdots \times s_n^{\sigma_n}$. Also define a vector generating function $F(t,s) = (F^1(t,s), \cdots, F^n(t,s))^T$. It follows from [1, p.119, theorem 3] and from assumption (7) that $F(t,s)$ satisfies the system of equations

$$\partial F(t,s)/\partial t = f(F(t,s)) \tag{1}$$

with the initial conditions $F(0,s) = s$. Here $f(s) = (f^1(s), \ldots, f^n(s))$ where $f_i(s) = \gamma_i - (\gamma_i + \beta_i)s_i + \beta_i s_i s_{i+1}$ (we put $\beta_n = 0$). It is impossible to find an explicit solution of the system (1) for all initial conditions and all values of parameters. Nevertheless, behaviour of the average number of sign meanings $M(t)$ at the moment t is described by the system of linear differential equations. It is possible to obtain the following formula for $M(t) = L_1 p_1(t) + G_1 L_2 p_2(t) + \ldots + G_1 \ldots G_{n-1} L_n p_n(t)$ where $p_i(t)$ for $t \geq 0$ is a density for the sum of i exponentially distributed independent random variables of means L_1, …, L_i.

Theorem 1. Under assumptions (1)–(7) we have only two qualitatively different kinds of behaviour for $M(t)$ when $t \geq 0$:

1. If $G_1 > 1$ then there exists a unique maximum at $t^* > 0$: $M(t^*) > M(t)$ for all $t \in [0, \infty] \setminus \{t^*\}$. Also $M'(t) \geq 0$ for all $t \in [0, t^*]$ and $M'(t) \leq 0$ for all $t \in [t^*, \infty]$.

2. If $G_1 < 1$ then $M'(t) \leq 0$ for all $t \geq 0$ and $M(t)$ reaches its global maximum at $t^* = 0$: $M(0) = 1$.

5 Regularities for the dynamics of a sign's sense volume, frequency of use, length, etc.

These regularities are deduced on the basis of conclusions made earlier on the polysemy quantitative dynamics and some other "qualitative" assumptions (see "Basic assumptions").

6 Arriving at the conclusion on general shape of momentary word polysemy, frequency of use, length, etc. distributions in language.

For deducing the general shape of momentary distribution in language for these features we assumed that some independent source generates signs for a language with some constant rate β_0. It is possible to show that in time this system arrives at some stationary state and to find its numerical characteristics.

Further details on this point, analytical deriving of other dependences, as well as presenting of some empirical data for testing the deduced form of polysemy distributions of lexemic signs (words) in various types of dictionaries of languages of various types — Russian, English, Chinese, Vietnamese, Mongolian, Hungarian, Estonian, Turkmen, Turkic, Tartar, Azerbaijan, etc. — will be made in the extended version of this paper.

References

Sevast'janov B.A. (1976) Vetvyashchiesya protsessy. (Russian) [Branching processes] Izdat. "Nauka", Moscow.