ITHEA

International Journal

INFORMATION TECHNOLOGIES & KNOWLEDGE

2010 Volume 4 Number 3

International Journal
INFORMATION TECHNOLOGIES & KNOWLEDGE
Volume 4 / 2010, Number 3

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" (IJ ITK)
is official publisher of the scientific papers of the members of
the ITHEA International Scientific Society

IJ ITK rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITK as well as the **subscription fees** are given on ***www.foibg.com***.

Responsibility for papers published in IJ ITK belongs to authors.

General Sponsor of IJ ITK is the **Consortium FOI Bulgaria** (www.foibg.com).

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Vol.4, Number 3, 2010

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with:
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Institute of Mathematics and Informatics, BAS, Bulgaria,
Universidad Politécnica de Madrid, Spain.

Publisher **ITHEA**
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, www.foibg.com, e-mail: info@foibg.com
Printed in Bulgaria

ISSN 1313-0455 (printed) **ISSN 1313-048X (online)** **ISSN 1313-0501 (CD/DVD)**

GRAMMATICAL PRIMING DOES FACILITATE VISUAL WORD NAMING, AT LEAST IN SERBIAN

Dejan Lalović

Abstract: Starting from the seminal work in 1980s to more recent findings, literature review suggests grammatical priming to be an elusive fenomenon, reliably obtained mostly in a lexical decision task and only rarely in naming task. Prevalent conclusion derived from the aforementioned fact suggests the effects of grammatical priming to be of less importance for online word processing as reflected by naming. However, this goes against intuitive notion of grammatical information being especially valuable in processing richly-inflected, free-word ordered language such as Serbian. The conclusion was challenged in a naming task in which prepositions and personal pronouns were employed to prime target nouns and verbs. We also tested the effect of prime-target asynchrony at 600ms and 250ms intervals, as the variable is known to invertly influence effects of language priming. Delayed naming condition was used to provide a purer estimate of target processing time afforded at the two asynchrony intervals in online naming. Analyses suggest effects of grammatical priming to be both substantial and robust. The facilitation of 22 ms (25 ms at 600 ms asynchrony, 20 ms at 250 ms asynchrony) provided by grammatical information was roughly twice as large as obtained in comparable studies in English. The facilitation effect was not qualified by interaction with SOA and therefore should not be attributed to some major strategic process associated with the longer SOA. We conclude grammatical priming in naming to be possible, at least in case of richly-inflected, free word-ordered language, and more than one word class primed. Online-delayed average latencies difference indicated slightly wider time window for target processing at the shorter asynchrony. The fact requires caution in grammatical priming effects loci interpretation.

Keywords: grammatical priming; word naming.

ACM Classification Keywords: I.2 Artificial Intelligence; I.2.7 Natural Language Processing – Language parsing and understanding.

Introduction

The research presented re-examines effects of grammatical information provided by single-word grammatical priming in word naming. The evidence of robust effects of such priming in Serbian word naming will be laid out along with certain considerations and further steps in the fenomenon investigation.

Grammatical priming could be considered a special case of syntactic priming, in which single, usually but not necessarily, function word acts as a prime to constrain some open-class target word grammatical properties. For instance, in English the definite article *the* constrains following target word class to nouns (e. g. *the law*, but not

the judged); in Serbian prepositon *kroz* (through) constrains both the target word class (noun) and the case (accusative). Starting from the seminal work of 1980s to more recent research, literature review suggests grammatical priming to be reliably obtained mostly in a lexical decision task and rarely in naming task. Goodman et al. were the first to demonstrate grammatical priming in lexical decision, employing single function-word primes [Goodman et al., 1981]. They used words with the function similar to that of articles and pronouns in English to prime either congruously *(it tied, no bread)* or incongruously *(no tied, it bread)* target verbs and nouns. The effects of such grammatical priming were 15-19ms in magnitude, i.e. more than twice weaker than the effects of semantic priming obtained in the same study. Seidenberg was able to replicate Goodman's findings using the same stimuli only in a lexical decision task. Grammatical priming had only marginal effects in naming task [Seidenberg et al, 1984]. Seidenberg et al. discussed the finding in terms on pre- and postlexical loci of word processing, stipulating that grammatical priming operates only at the level of postlexical processing required by lexical decision. West and Stanovich [1986] further validated the notion of grammatical priming effects restricted to postlexical processing. In their study syntactic priming was employed on the same materials in both lexical decision and naming task. Materials were taken from Wright and Garret's study [1984]. West and Stanovich used sentence fragments ending either with modal verb (e. g. *could, would*) or with preposition (e. g. *of*, *through*) to prime target verbs and nouns either congruously or incongruously. In series of experiments they varied procedural variables (reading priming context silently or overtly, pace of context presentation, materials employed, etc.) to reach the conclusion that syntactical priming effects are robust and comparable in magnitude in both word processing paradigms. Stronger priming effects on naming were obtained in Experiments 1 and 2 where average naming latencies were longer than in Experiments 3 and 4. West and Stanovich were prone to conclude naming task not to be immune to postlexical processing in cases of prolonged naming latencies, thus assigning syntactical priming effects in naming to that type of processes. Establishing the loci of language priming effects was the purpose of Sereno's study, in which graphemic, associative and grammatical priming in backward masked primed lexical decision and naming was employed [Sereno, 1991]. A single set of procedural variables and an extremely short prime exposition of 60ms was employed in the three types of language priming. The short exposition was introduced to limit possible strategic influences not directly associated with online word processing, such as generating expectancies about the target, etc. Sereno showed that grammatically congruous priming (verbs with modal verbs, nouns with determiners or possessive pronouns) facilitated only lexical decision by 28ms, while there was no analogue effect on target word naming. Such an outcome further favors the notion of grammatical priming effects being postlexical in nature. Bowey in her more recent study examined single-word grammatical priming in children word naming [Bowey, 1996]. While there is a general agreement of semantic context effects to be greater for poor readers than for good readers [Perfetti et al., 1979; Simpson et al., 1983] and for children than for adults [Simpson & Lorsbach, 1983; West & Stanovich, 1978], little is known of grammatical context effects in those groups. Subjects in Bowey's study were fourth-grade children (average CA 9 years 3,75 months), target stimuli were uninflected and inflected nouns and verbs primed by pronouns and numerals (nouns) or by modal verbs (verbs) in grammatically constrained condition. In the grammatically

unconstrained condition target words of both types were primed by conjectures *and/or*, which did not preclude neither target word class nor its form. Priming effects were evaluated at three different SOAs: 750ms (Experiment 1), 700ms and 400ms (Experiment 2). The facilitation obtained by priming in grammatically constrained condition was respectively 16ms, 13ms and 9ms, respectively in three stimulus onset asynchrony (SOA[1]) conditions. Difference of facilitation obtained under 700ms SOA condition and 400ms SOA condition proved not to be significant, leading to the conclusion of grammatical priming effects not likely to be attributed to some major strategic process associated with longer SOA. Finding that inflected and uninflected target word forms naming was equally facilitated by grammatical priming led to a tentative conclusion that the primary purpose of grammatical context is to constrain the target word class.

Initial research of grammatical priming in Serbian was published briefly following the paper of Goodman et al. [1981]. The research was guided by the idea of replicating and further exploring grammatical priming in language of rich inflectional morphology, with almost entirely free-word ordered syntax. In the first of two papers, Lukatela et al. [1982] demonstrated grammatical priming of inflected verb forms by congruent personal pronouns at SOA of 300ms (63ms effect) and at SOA of 800ms (96ms effect) in lexical decision task. The next study demonstrated similar effects of congruous priming of inflected noun forms by prepositions in which the prime exposition was paced by subjects [Lukatela et al., 1983] . In both studies facilitation by congruous priming was estimated with respect to the situation of incongruous priming of the same target stimuli (verbs primed by incongruous personal pronouns, nouns primed by incongruous prepositions). The hypothesis derived from the volume of research described was that grammatical priming in Serbian occurs at the level of inflexions – suffixes which specify open-class word forms thus specifying their grammatical properties. In research to follow the same group of researchers further explored the hypothesis outlined by employing properties of Serbian inflectional morphology. In the first of two studies of such type, Gurjanov et al. [1985a] primed target nouns by adjectives consistent with respect to the noun gender, number and case all marked with inflectional suffix (LEP*A* ŽEN*A* – beautiful woman).The task employed was lexical decision, again self paced as in Lukatela et al., 1983 [Gurjanov et al., 1985a]. Facilitation by congruous priming was measured against the condition of incongruous priming (primes and targets not matching in inflection, thus not matching in grammatical properties stated). Similar facilitation was observed when adjectives and adjectives looking like pseudowords with congruous or incongruous inflexions, primed target nouns. Such outcomes further favored the notion of grammatical priming acting upon a level of Serbian word inflexions. In the following study, possessive pronouns were used to prime target words in lexical decision task [Gurjanov et al., 1985b]. Facilitation by congruent priming (MOJA LOPTA – mine ball) was observed against the situation of priming the same targets with a look a like but nonsense pseudopronouns (MEJA LOPTA). The finding of nouns primed by congruent possessive pronouns was replicated in Lukatela et al. research [1987]. Further, it was observed that priming of target pseudonouns is also possible, as long as they share the same inflections with possessive pronoun primes. In a more recent research, Serbian feminine nouns in nominative

[1] Interval of prime exposition plus latency between prime offset and target onset

were primed by neutral visual prime *** and in the accusative case either by the same neutral visual prime or by congruent prepositions in a lexical decision task [Katz et al., 1995]. Congruent accusative priming yielded an effect of roughly 25ms facilitation as compared to a neutral situation defined by priming the same case with ***. However, the research of Katz et al. could not be regarded as a strictly grammatical priming study despite the materials employed, since the primes and targets were *simultaneously* visually presented.

The only study to directly compare effects of grammatical priming in lexical decision and naming task was the one by Carello et al. [1988] . In Experiment 1 effects of associative and grammatical priming of nouns by congruent and incongruent possessive pronouns in lexical decision and naming were contrasted. Carello et al. found both tasks to be sensitive to associative priming but only lexical decision to be sensitive to grammatical priming. Experiment 2 employed the same materials as it was employed in Lukatela et al's initial study in Serbian [Lukatela et al., 1982]. The effects Lukatela et al. [1982] obtained on identical materials in lexical decision were 63ms (300ms SOA) and 96ms (800ms SOA) were among the strongest grammatical priming effects. At the SOA of 600ms Carello et al. were not able to establish differences between naming latencies in situations where either target verbs or pseudoverbs were primed by congruous pronoun, incongruous pronoun or by pseudopronoun, respectively. Carello et al. [1988] concluded grammatical priming to be a phenomenon restricted to a lexical decision task. Naming, in their opinion, appears to be immune to automatic postlexical prime-target coherence check that influences lexical decision [Carello et al., 1988, p. 193].

Evidence on the effects of grammatical priming on word recognition is not restricted to data obtained in experimental paradigms employing strictly visual prime presentation. There is ample evidence of grammatical priming effects in cross-modal priming studies conducted in several languages other than English: Italian [Bentrovato et al., 1999; Bates et al., 1996; Bates et al., 1995], German [Hillert & Bates, 1996], Russian [Akhutina et al., 1999] and Chinese [Lu et al., 2000]. However, variations in methods employed and differences in the respective languages' syntax and morphology make such evidence rather suggestive than conclusive.

Prevalent conclusion derived mainly from studies in English of grammatical priming acting upon postlexical processing since it has been reliably observed only in lexical decision task and therefore being of less importance for online word processing, would seem premature in light of the evidence already presented. First, besides failed attempts, several studies indicated grammatical priming to affect visual word naming in English. Second and arguably more important, it seems premature to disregard the phenomenon of grammatical priming in word naming on account of mixed results obtained predominantly in one language. Grammatical properties of Serbian favor another scenario, we believe. The Serbian morphology, unlike the English morphology, is a highly inflected one. All open class words, as well as some types of closed class words, consist of a stem to which inflexional affix is appended to specify the word's grammatical attributes (e. g. noun's case, grammatical number and gender). Each inflexion specifies several possible word's thematic roles in a sentence (e. g. subject, object). At the same time, the Serbian sentence word order is almost entirely free. Unlike the English or German syntax, for instance, it means that a word's position within a sentence does not convey much information about its thematic role. Grammatical agreement – of a preposition with a noun, of an adjective with a noun, of a personal pronoun

with a verb, to mention but the few examples employed in the aforementioned research in Serbian – should therefore provide a powerful tool in thematic role of the single word disambiguating or in sentence parsing.

The aim of our research was to establish grammatical priming effects in Serbian word naming by employing what seems to be the crucial common feature of the studies which obtained such effects in naming. Following the implications of research in English succeeding in demonstrating grammatical priming in naming [e. g. Bowey, 1996; West & Stanovich, 1986], we primed two target word classes: nouns and verbs. Priming nouns with consistent prepositions and verbs with personal pronouns unequivocally constrained both the target word class and the form in the constrained condition. In the unconstrained condition, conjectures I/ILI (and/or) were employed. Such linguistic primes did not preclude neither the target word class, nor its form. Targets were chosen from a low frequency range, target frequency being suspected to be invertly related to priming effects [Bowey, 1996]. Two SOAs were employed: one of 600ms and the other of 250ms. The longer SOA could be considered a standard in the research in Serbian and was employed in the only grammatical priming in Serbian naming study [Carello et al., 1988]. Priming at the more than twice shorter SOA was introduced in order to test robustness of the phenomenon under scrutiny. Finally, the delayed naming condition was introduced to account for extra linguistic performance components in online primed naming and therefore to provide a more precise estimate of processing time afforded at two different online naming SOAs.

Experiment Subjects

Subjects were 60 psychology sophomores at the Faculty of Philosophy of Belgrade University, predominantly female. All of the subjects had Serbian as the mother tongue and vision either normal or corrected to normal. Participation in experiment was one way of fulfilling the course of Memory and Thinking requirements. All of the subjects had previous experience with visual word processing experiments. Each subject was assigned to one of two prime-target SOA interval conditions according to her/his order of appearance at the laboratory and was presented one of two sets of materials.

Materials

48 target nouns and 48 verbs were chosen from low frequency range in Serbian [Kostić, 1999]. One half of each target classes were 5-letter words and the other half 6-letter words of matching average frequency. All of the target nouns were female in gender, presented in two grammatical forms marked by the inflexional suffix: half with the inflexion –E (LOPTE, ball), the other half with the inflexion –U (LOPTU). All of the target verbs were in present tense, half of them third person singular (ZABODE, stabs), the other half third person plural (ZABODU). The target verbs in third person singular had the unique inflexional suffix –E and in third person plural unique inflexion –U which defined verb form. Words from both target classes had the same consonant-vowel structure: 5-letter

target words were either CCVCV or VCVCV, while all of 6-letter targets were CVCVCV in structure. Target nouns and verbs were balanced with respect to number of items of each of the structures described.

In the constrained condition, target nouns were primed with consistent prepositions: nouns with the inflexion –E with ZBOG (because) and BEZ (without) specifying target noun to be in genitive case; nouns with the inflexion –U with KROZ (through) or UZ (along) specifying target noun to be in accusative. In the same condition, target verbs were primed by consistent personal pronouns: present third person singular verbs with ON (he) or ONA (she); present third person plural with ONI (they, male) or ONE (they, female). In the unconstrained condition, both target nouns and verbs were primed by conjectures I/ILI. Example of the materials is presented in Table 1.

Table 1

Typical prime-target pairs as a function of target word type, and constraint

	Constrained Condition		*Unconstrained Condition*	
	Primes	*Targets*	*Primes*	*Targets*
	ZBOG/BEZ	IVICE/TARABE	I/ILI	KREDE/PALICE
	(because/without)	(edge/fence)	(and/or)	(chalk/bat)
Nouns				
	KROZ/UZ	SCENU/TERASU	I/ILI	GLINU/KOLIBU
	(through/along)	(scene/terrace)	(and/or)	(clay/cottage)
	ON/ONA	SNUJE/RUKUJE	I/ILI	GREBE/ZABODE
	(he/she)	(dreams/handles)	(and/or)	(scratches/stabs)
Verbs				
	ONI/ONE	TRUJU/DIRAJU	I/ILI	UKINU/DOVEDU
	(they m./they f.)	(poison/touch)	(and/or)	(eliminate/bring)

Two sets of materials were constructed. Half of the each of four target word types was randomly selected to be presented in constrained and unconstrained condition in the first set. In the second set of materials, targets that were in the first set presented in constrained condition were assigned to the unconstrained condition while the targets presented in the unconstrained condition in the first set were assigned to constrained condition. Within both of the sets, half of targets presented in grammatically unconstrained condition were primed by neutral prime I, while the other half was primed by neutral prime ILI. In both of sets in constrained condition half of randomly

selected target nouns with the inflection –E were primed by preposition ZBOG, the other half in the same condition with preposition BEZ; half of randomly selected target nouns with the inflection –U were primed by preposition KROZ, the other half in the same condition with preposition UZ. The same principle in set composing was applied to target verbs: half of randomly selected verbs present third person singular were primed with personal pronoun ON, the other half in the same condition with personal pronoun ONA; half of randomly selected verbs present third person plural were primed with personal pronoun ONI, the other half in the same condition with personal pronoun ONE. Each prime-target pair in both constrained and unconstrained condition constituted semantically acceptable syntagm.

Each set comprised an equal number of the four type prime target-pairs. Targets were arranged in a quasirandom order within the first list and primes were assigned to them as described above. Target order was retained in the second material set while the primes in the set were counterbalanced. Half of the subjects saw the first set and the other half of subjects saw the other set of stimuli at both of the SOAs. Each target was presented only once within a set, in the first set in constraining and in the second set in unconstraining grammatical context. Each subject therefore read each target only once, in one of two grammatical contexts. Both sets were presented to an equal number of subjects.

Subjects were given 16 practice trials before the set presenting, to adjust to the experimental procedure.

Procedure

Experiment was run by AT 486 PC connected to 14" CRT monitor. Stimuli exposition and naming latencies recording were controlled by SuperLab Pro 2.0 software. Naming latencies were collected with a Genius PC microphone with fixed stand. Subjects were reminded in the course of experiment to keep the distance from the microphone constant by holding chin above the line drawn on the table 30cm from microphone. Latencies were measured from the onset of the target word until subject's voice has reached predefined loudness threshold. All the stimuli were presented in capital Times New Roman Latin letters.

In online naming task, each trial started with warning signal sign ! which remained on the screen for 750ms. Following the warning signal, a prime was presented in the same line. In the 600ms SOA condition, prime exposition was 500ms followed by 100ms blank screen period; in 250ms SOA condition prime exposition was 150ms followed by 100ms blank screen. After the SOA expired, target was presented in the same line warning signal and prime were previously displayed. Target remained on the screen until subject vocalized and loudness reached threshold. Position of the prime's last letter and target's first letter was held constant, thus preserving the same prime-target distance in all of trials. Interval between two successive trials was 1s.

Subjects were tested individually, in a quiet room. Before experiment commencing, each subject read the instructions from the screen, afterwards to be briefly summarized by experimenter. Instructions equally stressed

importance of speed and naming accuracy. Subjects were warned that they will be asked at random to repeat a prime read silently, which happened on average after eight experimental trials. In case of prime missed or not correctly repeated, subjects were again politely asked to read primes carefully.

Cases of target mispronunciation and some sound other then subject's voice microphone (e. g. cough) triggering were hand coded by experimenter. Experimenter also noted cases of target read not loud sufficiently to trigger the microphone and cases of prime missed or not repeated correctly. Errors in target naming (mispronunciations) were analyzed in error analyses. All the other trials described were treated as technical errors and excluded from reaction time analyses. Mispronunciations and technical errors put together constituted spoiled trials.

Delayed naming task was conducted briefly after the online naming, in the same experimental session. In this task subjects named targets presented in online naming. Subjects were warned with the ! signal of 750ms duration before each trial started. All targets were displayed in center of the screen and remained there for 1500ms after which period they were put in brackets. Brackets were cue for subject to pronounce the word. Intertrial period was 1s. Instructions displayed on the screen and afterwards summarized by experimenter asked subjects to read word silently immediately after presented, to prepare pronunciation and to pronounce it only after the brackets around the word appeared. Errors were recorded as in online naming. It took on average 25 minutes for the whole experimental session to be completed.

Results

Latencies in excess of 1400ms and less than 300ms were excluded from reaction time (RT) analysis and added to spoiled trials. Data from 387 (6,7%) trials were discarded as spoiled trials. Out of these, 63 trials (1,1%) represented target naming errors (mispronunciations); 296 trials (5,1%) represented technical errors while naming latencies in 28 trials (0,5%) fell out of the RT range specified. Spoiled trials were treated as missing data in subject and item RT means calculations and were not replaced. Average online naming latencies are presented in Table 2.

Table 2

Target word online naming latencies (in milliseconds) and [SD] as a function of Word Class, SOA, and Constraint

	SOA 600 ms		*SOA 250 ms*	
	CONSTRAINED Mean [SD]	*UNCONSTRAINED* Mean [SD]	*CONSTRAINED* Mean [SD]	*UNCONSTRAINED* Mean [SD]
NOUNS	696 [82]	724 [96]	750 [97]	770 [102]
VERBS	682 [84]	703 [87]	743 [92]	761 [95]

Statistical analyses (by subjects and by stimuli) were conducted using univariate ANOVAs. The ANOVA for subjects included the following factors: SOA (between factor, levels: 600ms and 250 ms), Grammatical Constraint

(within factor, levels: constrained, unconstrained), and Word Class (within factor, levels: nouns, verbs). The ANOVA performed on the subjects online RTs indicated significant main effects of SOA [F(1, 58) = 5,47, p<0,05], Grammatical Constraint [F(1, 58) = 58,23, p<0,0001], and Word Class [F(1, 58) = 22,43, p<0,0001]. None of the interactions was significant. In the subject analysis main effects of SOA (within factor) was significant [F(1, 94) = 7,68, p<0,01], Grammatical Constraint (within factor) was significant [F(1, 94) = 65,63, p<0,0001)] and also was the between factor Word Class [F(1,94) = 34,28, p<0,0001)].

Main outcome of the online naming latencies analysis was that grammatically constrained targets were named 25ms faster than unconstrained targets at the SOA of 600ms, and that the 20ms effect of such facilitation was observed at the shorter SOA of 250ms. The facilitation effect was not qualified by SOA x Grammatical Constraint interaction in neither of analyses. Word Class effect of verbs being pronounced more rapidly than nouns is not of central interest, since not qualified by any of interactions.

Table 3 summarizes average error rates in online naming.

Table 3

Average error rates (in %) and [SD] as a function of Word Class, SOA, and Constraint

	SOA 600 ms		*SOA 250 ms*	
	CONSTRAINED Mean % [SD]	*UNCONSTRAINED* Mean % [SD]	*CONSTRAINED* Mean % [SD]	*UNCONSTRAINED* Mean % [SD]
NOUNS	0,6 [1,5]	0,7 [1,6]	1,8 [3,1]	0,4 [1,3]
VERBS	1,0 [1,9]	0,9 [2,1]	3,0 [3,6]	0,9 [1,8]

In the error analysis with the same factors as in ANOVA for average latencies, main effect of SOA [Subjects: F(1, 58) = 4,88, p<0,05; Stimuli: F(1,94) = 5,12, p<0,05], Word Class [Subjects: F(1, 58) = 12,58, p<0.001; Stimuli: F(1,94) = 14,56, p<0,001] and interaction of SOA x Word Class [Subjects: F(1, 58) = 12,86, p<0.001; Stimuli F(1,94) = 13,97, p<0,001] were significant. Again, SOA did not interact with Grammatical Constraint, but approached significance [Subjects: F(1, 58) = 3,06, p>0.09; Stimuli: F (1,94) = 4,11, p>0,07].

Overall error rate in the experiment (1,1%) could be considered fairly small. Main SOA effect and SOA x Grammatical Constraint interaction approaching significance are apparently due to more errors committed in the *constrained* condition at the SOA 250ms, as the Table 3 points out. SOA x Word Class interaction also has no important implications, since it stems from the same counterintuitive unconstrained condition advantage more pronounced in case of verb targets. Inspection of individual subjects data suggest such outcome could be ascribed to the data from five subjects who have committed significantly more errors (up to 9%) in the constrained condition. However, closer inspection of their naming latencies hinted no speed-accuracy trade off or any other anomaly in performance.

Finally, we present the delayed naming analyses. Average delayed naming latencies are presented in Table 4.

Table 4

Target word delayed naming latencies (in milliseconds) and [SD] as a function of Word Class, SOA, and Constraint

	SOA 600 ms		*SOA 250 ms*	
	CONSTRAINED Mean [SD]	*UNCONSTRAINED* Mean [SD]	*CONSTRAINED* Mean [SD]	*UNCONSTRAINED* Mean [SD]
NOUNS	578 [106]	571 [112]	571 [102]	569 [99]
VERBS	578 [112]	578 [101]	575 [107]	576 [103]

In delayed naming RT analyses parallel to those conducted on the online naming RTs and errors no significant effect was obtained. Average delayed naming latency at the 600ms SOA and at the 250ms SOA were 577ms and 573ms, respectively. Mispronunciations rate of 0,3% (spoiled trials 1,3%) rendered delayed naming error analysis uninformative.

Discussion and Conclusion

The obtained results show average 22ms facilitation of word naming by grammatical priming. Such effect can be considered substantial, being roughly equal to or larger than effects in grammatical priming studies with English adult readers but in which a lexical decision task was employed (15-19ms Goodman et al., 1981; 13ms Seidenberg et al., 1984). Bowey obtained weaker effects (9-16ms) in naming task, with less than proficient English readers (Bowey, 1996). Our study was conservative since only facilitative effects of grammatical priming were investigated. Unlike most priming studies conducted in Serbian, we have measured only facilitation provided by congruent priming with respect to a neutral situation. Studies in Serbian typically estimated facilitative effects of congruous grammatical priming with respect to a situation with incongruous grammatical priming [e. g. Lukatela et al., 1983; Lukatela et al., 1982], thus failing to distinguish facilitation effects from inhibition effects. In other studies in Serbian facilitation was estimated with respect to a neutral situation in which target words (and for that matter pseudowords) were primed by nonlinguistic stimuli such as *** [e. g. Katz et al., 1995; Carello et al., 1988], known to inflate effects of language priming when used in a baseline condition [deGroot et al., 1982]. In the neutral condition we employed linguistic primes as well as real target words across all situations.

Facilitation in our experiment can also be considered robust as obtained both at standard 600ms SOA (25ms effect) and at a brief SOA of 250ms (20ms effect). Bowey, for instance, obtained grammatical priming effects at the longer SOAs of 750ms, 700ms, and 400ms. The effect of grammatical priming was not qualified by interaction with SOA interval in our experiment, suggesting that manipulation with this procedural variable, i. e. long SOA,

was not likely to be the source of facilitation. Such outcome should be interpreted with caution, nevertheless. Certain authors hold the opinion that long prime-target SOAs are associated with attentional expectancies forming and with some strategic postlexical processes, like target word properties and prime agreement checking [e.g. Sereno, 1991; Seidenberg et al., 1984]. Short prime-target SOA apparently reduces chances for conscious expectancies or for any other strategic process to operate. However, average naming latency at the 600ms SOA in our experiment was 701ms while at the 250ms SOA it was 756ms. Average delayed naming latencies of subjects performing under the 600ms SOA were 577ms and at the 250ms SOA statistically equaled 573ms. Delayed naming was to provide control for performance factors in online naming not involved in pure visual word processing and for the voice key (microphone) sensitivity. When subtracted from online naming latencies, delayed naming latencies left time window of 125ms at the 600ms SOA and 183ms at the 250ms SOA for target processing. Almost 60ms longer time allowed for word processing at the shorter SOA leaves open the possibility of any strategic processes arguably acting at the longer SOA to take part also in naming with the shorter SOA interval. Notably, West and Stanovich [1986] obtained stronger syntactical priming effects in naming experiments with longer average latencies leading them to the conclusion naming task not to be immune to postlexical processing.

The crucial point of departure of our experiment from the only study of grammatically primed naming in Serbian [Carello et al., 1988] seems to be that we have primed two target word classes instead of only one. Carello et al. [1988] primed verb forms and failed to obtain grammatical priming effects in naming. A tentative conclusion from the fact would be that grammatical information provided by single word priming serves primarily to constrain a target word class, and only thereupon to specify target word properties within the class. A similar conclusion was put forward and corroborated in Bowey's study [1996]. This argument could be verified by contrasting outcomes of the experiment presented with two experiments in which target nouns and verbs we have employed would be grammatically primed in isolation. Priming of only one word class should diminish grammatical priming effects. The argument validity would reside solely on (priming) effects tests; experiments therefore should be planned with the a priori statistical power sufficient to detect or to reject presumably small effects. The next sensible step in grammatical priming investigation would be exploring lexical variables known to influence word naming. Target frequency and target length would constitute immediate candidates, both of them appearing to be invertly related to word recognition [West and Stanovich, 1986]. Thus, grammatical information should be of less importance in frequent and short words naming, and vice versa. However, main variable to influence grammatical priming in our opinion would be objectively registered prime-target coocurence frequency and subjectively assessed prime-target associative strength. If the variables could be proven to be of major influence in grammatical priming, that might add a contribution to the ongoing debate on the real nature of lexical priming. Namely, if semantic priming effects could be reduced to an associative priming mechanisms [e. g. Balota et al., 2006], the point of reducing grammatical priming to the same mechanisms would also seem viable in view of the type of evidence outlined above.

Acknowledgments

Research presented was partly supported by Republic of Serbia's Ministry for Science and Technological Development grant "Education Quality and Availability Improvement in Serbia's Modernization Processes" no. 47008 (2001-2014) awarded to the Institute for Pedagogical Research from Belgrade.

Bibliography

[Goodman et al., 1981] Goodman, G. O., McClelland, J. L., & Gibbs, R. W. (1981). The role of syntactic context in word recognition. Memory & Cognition, 9(6), 580-586.

[Seidenberg et al, 1984] Seidenberg, M. S., Waters, G. S., Sanders, M., & Langer, P. (1984). Pre- and postlexical loci of contextual effects on word recognition. Memory & Cognition, 12(4), 315-328.

[West and Stanovich,1986] West, R. F., & Stanovich, K. E. (1986). Robust effects of syntactic structure on visual word processing. Memory & Cognition, 14(2), 104-112.

[Wright and Garret, 1984] Wright, B., & Garret, M. (1984). Lexical decision in sentences: Effects of syntactic structure. Memory & Cognition, 12, 31-45.

[Sereno, 1991] Sereno, J. A. (1991). Graphemic, associative, and syntactic priming effects at a brief stimulus onset asynchrony in lexical decision and naming. Journal of Experimental Psychology: Learning, Memory, and Cognition, 3, 459-477.

[Bowey, 1996] Bowey, J. A. (1996). Grammatical priming of visual word recognition in fourth-grade children. The Quarterly Journal of Experimental Psychology, 4, 1005-1023.

[Perfetti et al., 1979] Perfetti, C. A., Goldman, S. R., & Hogaboam, T. W. (1979). Reading skill and the identification of words in discourse context. Memory and Cognition, 7, 273-282.

[Simpson et al., 1983] Simpson, G. B., Lorschbach, T. C., & Whitehouse, D. (1983). Encoding and contextual components of word recognition in good and poor readers. Journal of Experimental Child Psychology, 35, 161-171.

[Simpson & Lorsbach, 1983] Simpson, G. B., & Lorschbach, T. C. (1983). The development of automatic and conscious components of contextual facilitation. Child Development, 54, 760-772.

[West & Stanovich, 1978] West, R. F., and Stanovich, K. E. (1978). Automatic contextual facilitation in readers of three ages. Child Development, 49, 707-712.

[Lukatela et al. 1982] Lukatela, G., Morača, J., Stojnov, D., Savić, M. D., Katz, L., & Turvey, M. T. (1982). Grammatical priming effects between pronouns and inflected verb forms. Psychological Research, 44, 297-311.

[Lukatela et al., 1983] Lukatela, G., Kostić, A., Feldman, L. B., & Turvey, M. T. (1983). Grammatical priming of inflected nouns. Memory and Cognition, 1, 59-63.

[Gurjanov et al., 1985a] Gurjanov, M., Lukatela, G., Lukatela, K., Savić, M., & Turvey, M. T. (1985). Grammatical priming of inflected nouns by the gender of possessive adjectives. Journal of Experimental Psychology: Learning, Memory, and Cognition, 4, 692-701.

[Gurjanov et al., 1985b] Gurjanov, M., Lukatela, G., Moskovljević, J., Savić, M., & Turvey, M. T. (1985). Grammatical priming of inflected nouns by inflected adjectives. Cognition, 19, 55-71.

[Lukatela et al., 1987] Lukatela, G., Kostić, A., Todorović, D., Carello, C., & Turvey, M. T. (1987). Type and number of violations and the grammatical congruency effect in lexical decision. Psychological Research, 49, 37-43.

[Katz et al., 1995] Katz, L., Rexer, K., & Peter, M. (1995). Case morphology and thematic role in word recognition. In L. B. Feldman (Ed.), Morphological Aspects of Language Processing (pp. 79-107). Hillsdale, NJ: Lawrence Erlbaum Associates.

[Carello et al. 1988] Carello, C., Lukatela, G., & Turvey, M. T. (1988). Rapid naming is affected by association but not by syntax. Memory and Cognition, 3, 187-195.

[Bentrovato et al., 1999] Bentrovato, S., Devescovi, A., D'Amico, S., & Bates, E. (1999). Effect of grammatical gender and semantic context on lexical access in Italian. Journal of Psycholinguistic Research, 6, 677-693.

[Bates et al., 1996] Bates, E., Devescovi, A., Hernandez, A., & Pizzamiglio, L. (1996). Gender priming in Italian. Perception & Psychophysics, 7, 992-1004.

[Bates et al., 1995] Bates, E., Devescovi, A., Pizzamiglio, L., D'Amico, S., & Hernandez, A. (1995). Gender and lexical access in Italian. Perception & Psychophysics, 6, 847-862.

[Hillert and Bates, 1996] Hillert, D., & Bates, E. (1996). Morphological constraints to lexical access: Gender priming in German. (Techical Report 9601). University of California, San Diego.

[Akhutina et al., 1999] Akhutina, T., Kurgansky, A., Polinsky, M., & Bates, E. (1999). Processing of grammatical gender in a three-gender system: experimental evidence from Russian. Journal of Psycholinguistic Research, 6, 695-713.

[Lu et al., 2000] Lu, C.-C., Bates, E., Hung, D., Tzeng, O., Hsu, J., Tsai, C.-H., & Roe, K. (2000). Syntactic priming of nouns and verbs in Chinese. (Technical Report CND-0001). University of California, San Diego.

[Kostić, 1999] Kostić, Đ. (1999). Korpus srpskog jezika: Frekvencijski rečnik srpskog jezika. Beograd: Institut za eksperimentalnu fonetiku i patologiju govora i Laboratorija za eksperimentalnu psihologiju. (Serbian Corpus: Frequency Dictionary. Belgrade: Institute for experimental phonetics and speec pathology and Laboratory for Experimental Psychology.)

[de Groot et al., 1982] de Groot, A. M. B., Thomassen, A. J. W. M., & Hudson, P. T. W. (1982). Associative facilitation of word recognition as measured from neutral prime. Memory and Cognition, 4, 358-370.

[Balota et al., 2006] Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual Word Recognition: A Journey from Features to Meaning (A Travel Update). In M. J. Traxler and M. A. Gernschbacher (Eds.), Handbook of Psycholinguistics 2nd Edition (pp. 285-375). New York: Academic Press.

Author's Information

Dejan Lalović *– Department of Psychology, Faculty of Philosophy, University of Belgrade; Čika Ljubina 18-20 St. 11000 Belgrade, FR Serbia; e-mail:* dlalovic@f.bg.ac.rs

Major Fields of Scientific Interest: Working Memory; Psycholinguistics; Memory and Cognition in Mental Disorders.

SPAM AND PHISHING DETECTION IN VARIOUS LANGUAGES

Liana Ermakova

Abstract*: The majority of existing spam filtering techniques suffers from several serious disadvantages. Some of them provide many false positives. The others are suitable only for email filtering and may not be used in IM and social networks. Therefore content methods seem to be more efficient. One of them is based on signature retrieval. However it is not change resistant. There are enhancements (e.g. checksums) but they are extremely time and resource consuming. That is why the main objective of this research is to develop a transforming message detection method. To this end we have compared spam in various languages, namely English, French, Russian and Italian. For each language the number of examined messages including spam and notspam was about 1000. 135 quantitative features have been retrieved. Almost all these features do not depend on the language. They underlie the first step of the algorithm based on support vector machine. The next stage is to test the obtained results applying trigram approach. Proposed phishing detection technique is also based on SVM. Quantitative characteristics, message structure and key words are used as features. The obtaining results indicate the efficiency of the suggested approach.*

Keywords*: spam, corpus linguistics, phishing, filtering, text categorization.*

ACM Classification Keywords*: I.2.7 Text analysis*

Introduction

Kaspersky Lab defines spam in the following way:

Spam is unsolicited anonymous mass email [Kaspersky Lab, 2010].

According to Kaspersky lab, in the last quarter of 2010 spam made up 77.1% of total email traffic [Наместникова, 2011]. It should also be mentioned that Russian spam became more carefully designed: more spam messages have the HTML format [Kaspersky Lab, 2010]. Nowadays spam concerns not only email, but also social networks, instant messaging (IM) and other systems. Traditional approaches such as blacklisting and message header analysis are efficient enough for email filtering. Though, they fail to deal with spam in social networks, IM and forums. In this case content and link analyses seem to be more effective. Moreover the last two ones may be applied to identify phishing.

State-of-the-arte

Spam appeared in the nineties of the XX century. Firstly, spam was sent from proper spammers' addresses. The earliest messages were similar. That spam is easy to filter. Content analysis development forced spam to evolve. All messages became different. One of the ways to do it is to add an address to the beginning of a letter (e.g. adding «Hello, joe!» to the message to joe@user.com). The trick may be detected by applying fuzzy signature or statistical learning filters (like Bayesian filtering). A message may begin or end with an extract from classical literature or a sequence of random words. HTML message may contain an unreadable text (e.g. printed in very small font or the same color as the background). These additions provide obstacles to fuzzy signature and statistical filters. In response new techniques appear such as quotation searching and detailed HTML parsing. Usually it is possible to detect spammer's trick it-self and classify a message as spam without detailed content analysis. An advertisement may be sent as a picture. Therefore image analysis techniques which enable to retrieve a text from a picture are used.

Transforming messages are messages which have the same meaning but different forms. Every message looks like a connected text. Only if one has a number of these letters it is possible to establish a paraphrasing fact.

Nowadays the major part of junk emails is delivered from compromised user machines. The most widely used tricks are transforming messages, spam sent as a graphic attachment and unreadable text addition. And not all spam filters can deal with them [Kaspersky Lab, 2009].

Yandex divides spam detections methods into two categories:

- Techniques based on text samples (it is difficult to make them and to keep them up to date);
- Manual analysis and email monitoring (e.g. signature approach) [Yandex, 2010].

Yandex uses, inter alia, white listing [Yandex, 2010]. This approach suffers from some serious disadvantages. In the systems with authorization mechanism it is not so easy to send a message to a user for the first time. Moreover, the practice indicates that white listing is not efficient in IM (e.g. qip, icq) and social networks (e.g. ВКонтакте, Facebook) as far as there the lager half of spam is distributed from the accounts of authorized people. Some researchers believe that spam may be filtered only by end user [Сегалович, 2010]. According to another survey conducted by Yandex, 40% of the respondents have difficulties in distinguishing spam from legal mail [Kaspersky Lab, 2009].

Today the improvement of signature methods seems to be crucial. There are two basic approaches:

- Syntactical (i.e. operating with word chains);
- Lexical (i.e. operating with dictionary) (e.g. key words) [Yandex, 2010].

In current syntactical methods based on shingles (i.e. contiguous subsequences of tokens in a document) [Broder, 2003] [Manber, 1994], for each shingle a check sum is computed and then a random sample is constructed from this set. Shingles make it possible to find similar texts rather reliably. However, real-world

problems, such as spam filtering, require too many shingles and consequently too many resources in order to cluster messages [Yandex, 2010].

The major drawback of every lexical method is that it may be applied only to a single language.

Peculiarities of Spam in Various Languages

Spam classification may be made in terms of two criteria: by structure and by subject. Spam may be divided by structure into three types:

- Spam disguising as legal mass mailing;
- Spam disguising as a personal message;
- Advertising spam.

Regardless of the language, advertisement is spam dominating subject, especially medicine, tourism and education offers. English courses are very popular in non-English-speaking countries. Other subjects such as cheap software and pornography are common for various countries.

Advertising spam disguises as legal mass mailing and contains many links (especially French spam) and words related to a commerce sector. It often begins with an exclamatory or interrogative sentence. Bulleted and numbered lists are also common features of spam in various languages. Nevertheless these features may not be used for spam filtering since they occur in legal mass mails.

Another popular subject is easy money (Internet casino, lottery and so on). Sometimes it is related to phishing and identity theft as well as Nigerian scam. The latter resembles personal mail and is difficult to be filtered. Nigerian scam in French is designed according to the rules of business correspondence. However official letters usually contain an expression «à l'attention de» with a position and/or a name, while in spam one can see «à votre attention». There are a lot of email addresses in business correspondence as well as in phishing. The fraud is that a user may respond to a spam message. In this case the spammer will know that the email is active. The share of spam disguising personal messages is comparatively small. However it is necessary to take them seriously because legal messages can be lost.

French spam is more carefully designed than English and especially Russian ones. Usually it has HTML format therefore there are phrases like "Si ce mailling ne s'affiche pas correctement". Sometimes spammers suggest unsubscribing ("Cliquez ici pour ne plus recevoir nos emails"). If a person clicks on this link the spammer will know that this e-mail address is active and as a result the person will receive more spam or even download a virus. Sometimes spammers "explain" why people receive spam ("Vous êtes inscrit sur", "You are receiving this message because"). Due to perception peculiarity verb forms such as imperative, future simple and present are widespread in spam unlike solicited messages. Direct Impératif is not enough polite. Spammers try to control

readers and that is why imperative usually occurs in the aim of a junk mail («push the button now», «achetez maintenant»). An action in indicative mood is thought as a real one.

Many Anglicisms can be found in French and Russian spam. French spam contains less pronouns and possessive determinative than legal messages. There is not such a tendency in the Russian and English languages.

All types of spam appeal to feelings (curiosity, covetousness, laziness, credulity, boredom etc.). Spam features may appear according to subject, structure or aim of a message.

Methods of Message Transforming

Transliteration is often used in Russian spam. Besides, there are a lot of deliberate word distortions (e.g. unnecessary symbols, deliberate misprints, Latin letters in Cyrillic text etc.). However these features do not definitely indicates spam. Sometimes transliteration is used by emigrants and travellers for lack of Russian keyboard layout. Encoding problems may appear. People often apply different transliteration rules. In this case a human being may easily read a message but it is difficult to perform an automated analysis.

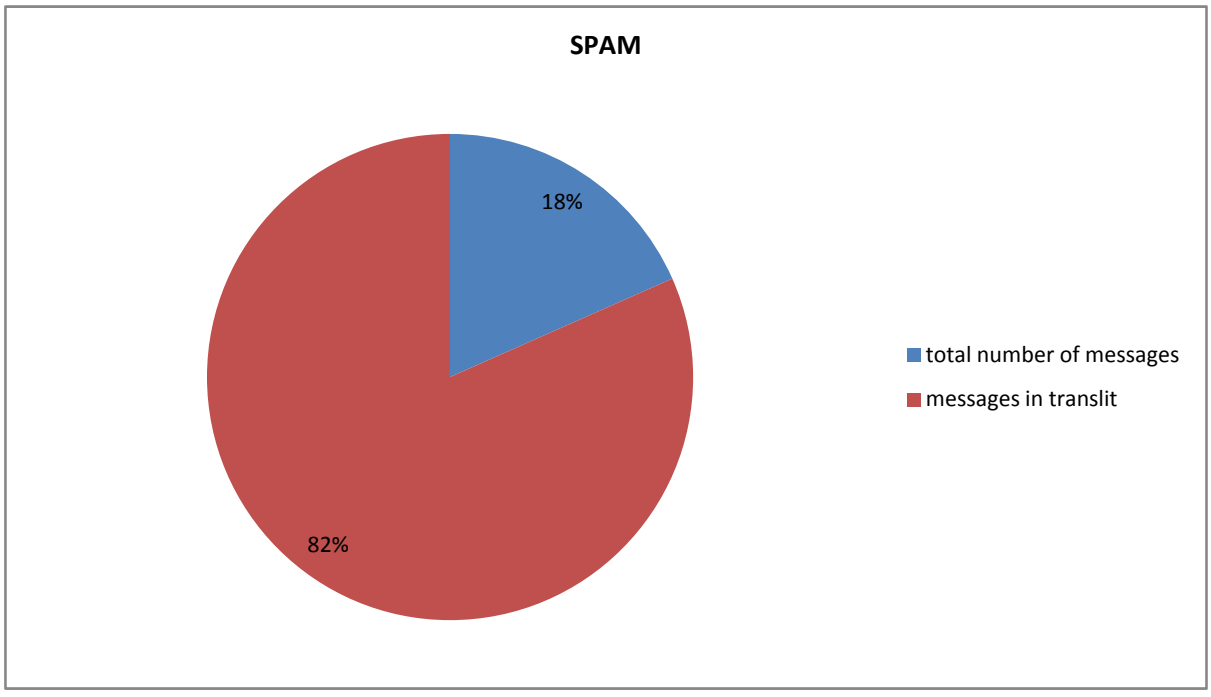

Fig. 1. Share of letter written in transliteration in spam

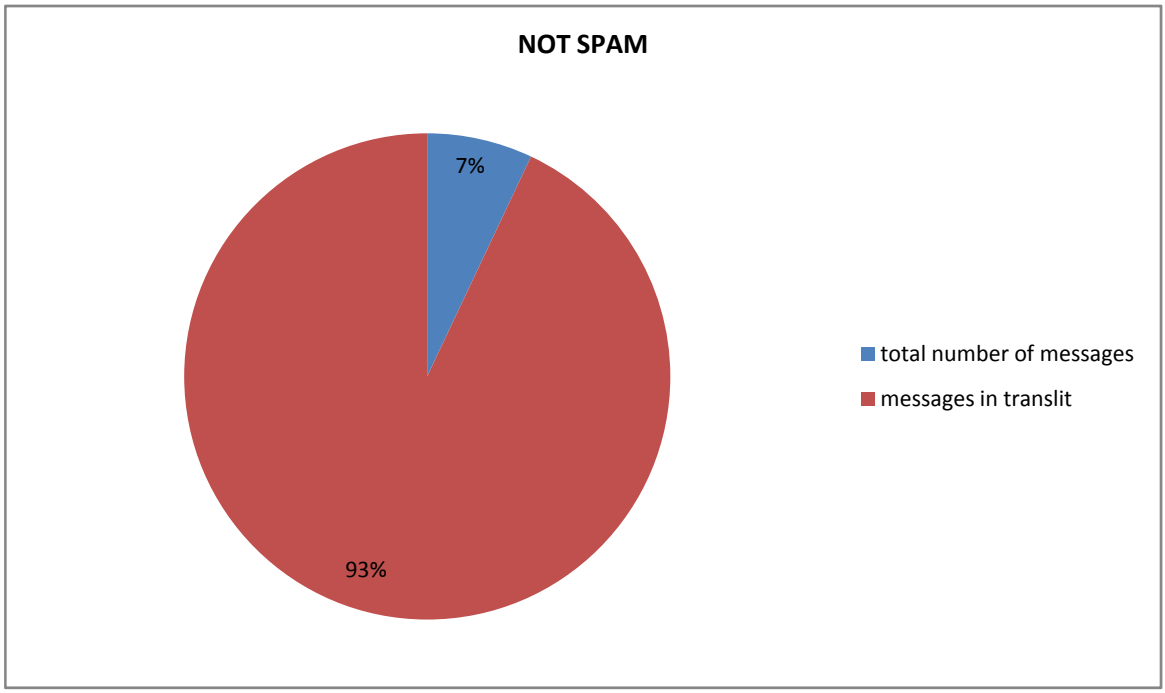

Fig. 2. Share of letter written in transliteration in legal messages

Spam	Not spam
pRODAVA email BAZ pRODAVA BAZ email ADRESOW (ADRESA DLQ email RASSYLOK) eSLI wY OBLADAETE SOBSTWENNYMI INSTRUMENTAMI PROWEDENIQ email RASSYLOK, TO DLQ wAS MY MOVEM PREDLOVITX BAZY DANNYH SOBSTWENNOGO SBORA. <…> cENA ZA 1 MLN. - 50$ cENA ZA WS@ BAZU - 500$ <…>PO L@BYM WOPROSAM: tELEFON:	Privet , zolotze. Nakonez-to posylayu tebe fotki. Ya vybrala nemnozhko bolshe , chto-by ty vybrala kakie hochesh i posmeyalas nemnozhko. Ya kogda smotrela, u menya srazu podnyalos nastroenie. Vse- taki my klassno s toboj syezdili v Ust- Kachku. Esli hochesh, ya tebe vse ostalnye tozhe pereshlu. Pishu tebe iz doma pervyj raz. Ladno, pobezhala delat chto- nibud. A - to zeloe utro za kompiuterom sizhu. Lublu, zeluyu. Mame i koshkam privet!

Here are some examples of transforming messages written in transliteration.

sWEVIE email BAZY pRODAVA BAZ email ADRESOW (ADRESA DLQ email RASSYLOK) <...>	В начале года всегда возникает потребность в "свежих" выписках ЕГРЮЛ и справках Госкомстата. Предлагаем Вам: получение выписки ЕГРЮЛ за 1,2ОО рублей справки Госкомстата за 1 200 руб. заказ выписки ЕГРЮЛ + справки Госкомстата составит всего 2.0ОО рублей Доставка курьером, оплата по факту. Контактная информация + 7495 222+07.68
sWEVIE email BAZY pRODAVA BAZ email ADRESOW (ADRESA DLQ email RASSYLOK) <...>	В начале года всегда возникает необходимость в "свежих" выписках ЕГРЮЛ и справках Госкомстата. Мы предлагаем Вам: получение выписки ЕГРЮЛ за 1 2ОО рублей справки Госкомстата за 1 тыс. 200 р. заказ выписки ЕГРЮЛ + справки Госкомстата составит всего 2 тыс. 0ОО руб-й. Доставка курьером, оплата по факту. Телефон: + 7495 222_07;68
aDRESA DLQ email RASSYLOK pRODAVA BAZ email ADRESOW (ADRESA DLQ email RASSYLOK) <...>	В начале года всегда возникает потребность в "свежих" выписках ЕГРЮЛ и справках Госкомстата. Мы предлагаем Вам: получение выписки ЕГРЮЛ за 1 тыс. 200 руб-й справки Госкомстата за 1 200 рублей. заказ выписки ЕГРЮЛ + справки Госкомстата составит всего 2,000 р. Доставка курьером, оплата по факту. Контакты + 7(495) 222-07-68

In Russian spam one can find a lot of "spammers' tricks":

- Substitution of letters by digits and vice versa (4-ч, 0-о, 3-з, 1-l);
- Substitution of Cyrillic symbols by similar Latin letters (k-к, a-а, H –Н и т.д.);
- Unnecessary symbols and blanks («Вы хотите ве рнуть вашего любимо го челове ка навсегда и полность ю избавиться от измен?»);
- Interchanging of different symbols (e.g., in telephone number).It is important to mention another transformation technique, namely synonymous expressions (sWEVIE email BAZY = sWEVIE email BAZY = aDRESA DLQ email RASSYLOK, Предлагаем Вам = Мы предлагаем Вам, необходимость= потребность).

It happens that only an address or a link transforms:

<...> La preghiamo di rispondere solo alla mia personale e-mail:khhaykanush@yahoo.com Tua amica Haykanush.
<...>La preghiamo di rispondere solo alla mia personale e-mail:haykanusharm@yahoo.com Tua amica Haykanush.
<...>La preghiamo di rispondere solo alla mia personale e-mail:khaykanush@yahoo.com Tua amica Haykanush.

Medicine advertisement is the most changeable. Both a subject and a text transform. They may even substitute each other. Usually all links are different (they are automatically created in free hosts). Meanwhile sense is the same.

Subject	Text
Desire to impress and please your lover tonight	The only bluepill you need to get bigger python. http://wanzulkifli.com/c6ave6lc.html
Gain in size and win your wife's addiction	Desire to act like a pornstar? Bang a magicpilule! http://bpyasociados.com.ar/9vh6w3lf.html
Wish to act like a porn-director Nail a blu colored med!	0% amorous failure risk http://mikloswowmobile.com/uaagzeib.html
Dream to act like a porn-director Bang a blu colored pil!	Long manliness is great http://antalyagunlugu.com/d4zz8qan.html

The same can be said about casino. It should be noticed that French and English spam is more intricate than Russian and Italian one; especially it concerns such areas as casino, medicine, stock market games, porno and software. In Spanish there are almost no transformations.

Subject	Text
Comme Faire _200 de _20 - nous APPRENDRONS	Bonne journee Jessikaparsons, {http://yxaqih983.o-f.com/kerizev.html} Accueillez la fortune dans votre vie avec de grandes opportunitйs de gagner, avec l'assurance que vos informations personnelles sont protйgйes et vos gains seront payйs rapidement. Une demi-heure et Ъ200 dans ta poche
Gagner _100 pour une demi-heure c'est rйel	Du jour reussi Shirley_patel, {http://gamingworldshop.ru} Il y a de grandes promotions auxquelles vous pouvez participer et qui vous promettent encore plus de plaisirs et de faзons de gagner. Faire Ъ100 pour une demi-heure - Apprendre?
Faire -100 pour une demi-heure - Apprendre	Bonne journee Nvshamshik, {http://beluwulod.maddsites.com/abimogek.html} Il y a de grandes promotions auxquelles vous pouvez participer et qui vous promettent encore plus de plaisirs et de faГons de gagner. Gagner -100 pour une demi-heure c'est rИel
Jouer ici, c'est le bonheur ! Telechargez maintenant	{http://opakypiwel.dreamstation.com/jededila.html} On ne peut pas faire plus simple, il suffit de vous inscrire, de faire un versement et vous recevez un fantastique bonus de bienvenue - alors foncez et gagnez ! La meilleure selection de jeu sur internet ! Jouez ici
Jouez plus, gagnez plus	Salut Shea.swan Des options bancaires sыres qui conviendront a tous sont disponibles. Relaxez-vous et soyez certains que vos informations confidentielles sont sйcurisйes et ne seront p& #97;s divulguйes. {http://durl.me/554k6}Comment aimeriez-vous commencer au mieux dans le jeu en ligne avec 1,200 Gratuits? Ils sont dйja a vous, rйclamez-les, jouez et gagnez!

Trigrams in Transforming Message Detection

There are many approaches to find the distance between two documents (e.g. Jaccard coefficient, Hamming distance, edit distance) [Chakrabarti, 2003]. In this research we have used trigram distance.

Traditionally trigrams are used in problems of plagiarism detection [Coulthard, 2004] [Halteren, 2004] and language and encoding identification [Sotnik, 2006] [Cavnar, 1994]. Another group of affiliation methods is based on quantitative text characteristics [Mesheryakov R., 2005] [В.П.Фоменко, 1983] [Рахимова, 2005]. Firstly quantitative features were used in Flesch index and Flesch-Kincaid Index [Галяшина, 2003]. Within the bound of this work these two approaches have been combined. We have used 135 quantitative text features such as share

of content and function words, share of sentences, paragraphs and words of specified length, share of various parts of speech (POS), punctuation marks, co-occurrence of POS etc. Trigram method was modified. Firstly, we have considered as a gram a word and not a symbol. We have examined the sequence of 3 elements in order to determine POS using Zalizniak's grammar dictionary.

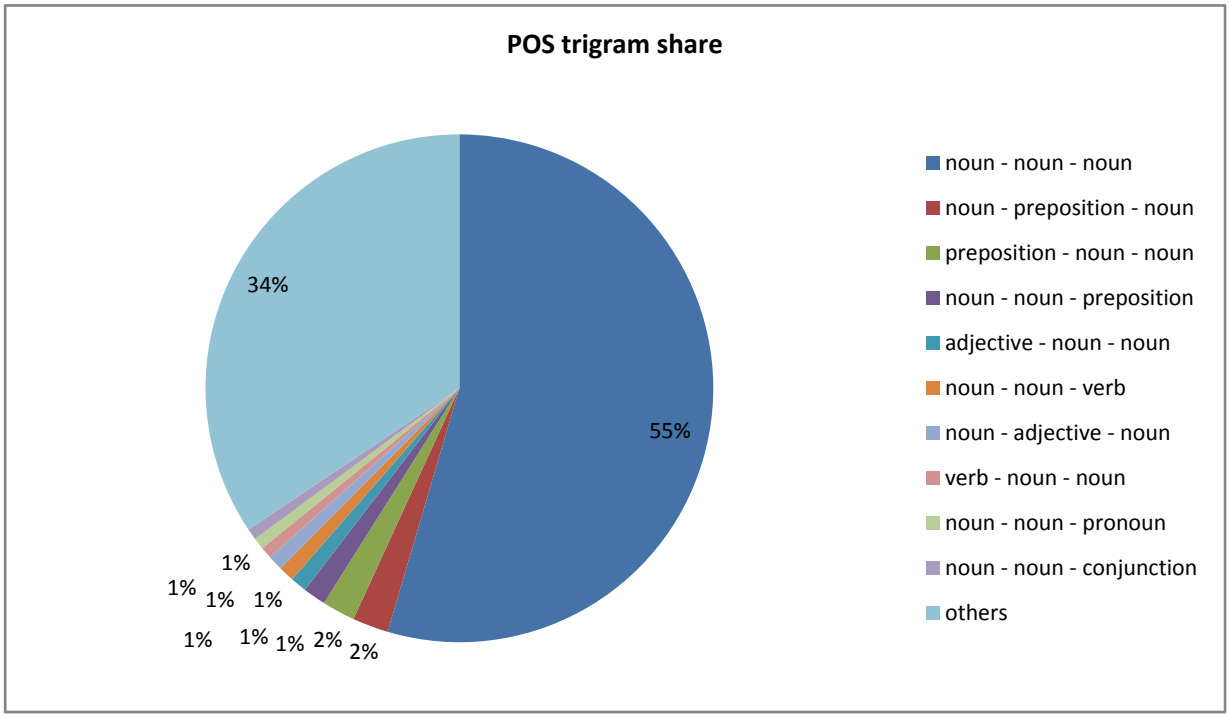

Fig. 3. Share of various POS sequences

Secondly, we have computed similarity of two messages:

similarity=2*NumberOfMatches/(NumberOfTrigramsIn_1_text+NumberOfTrigramsIn_2_text)

This quantity is not normalized. Similarity of Russian and Italian transforming messages is extremely high. Moreover, it varies slightly. Similarity of English and French letters is much smaller and has a large scatter (Fig. 4 - Fig. 9).

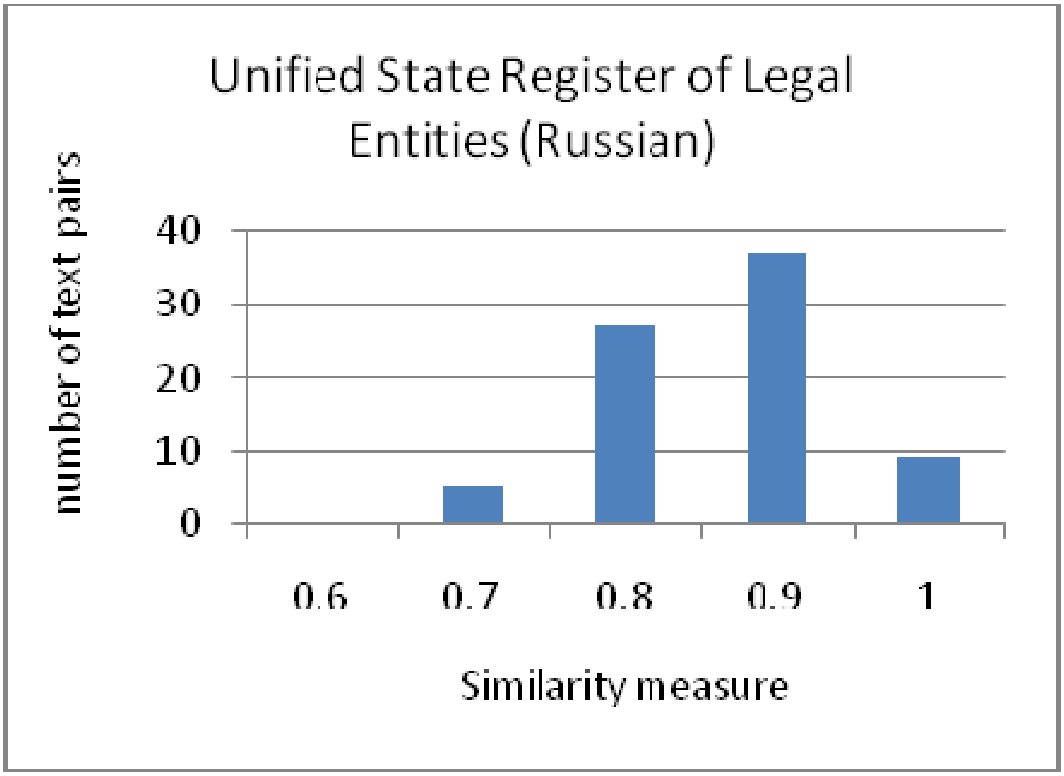

Fig. 3. Trigram similarity measure of "ЕГРЮЛ" mails

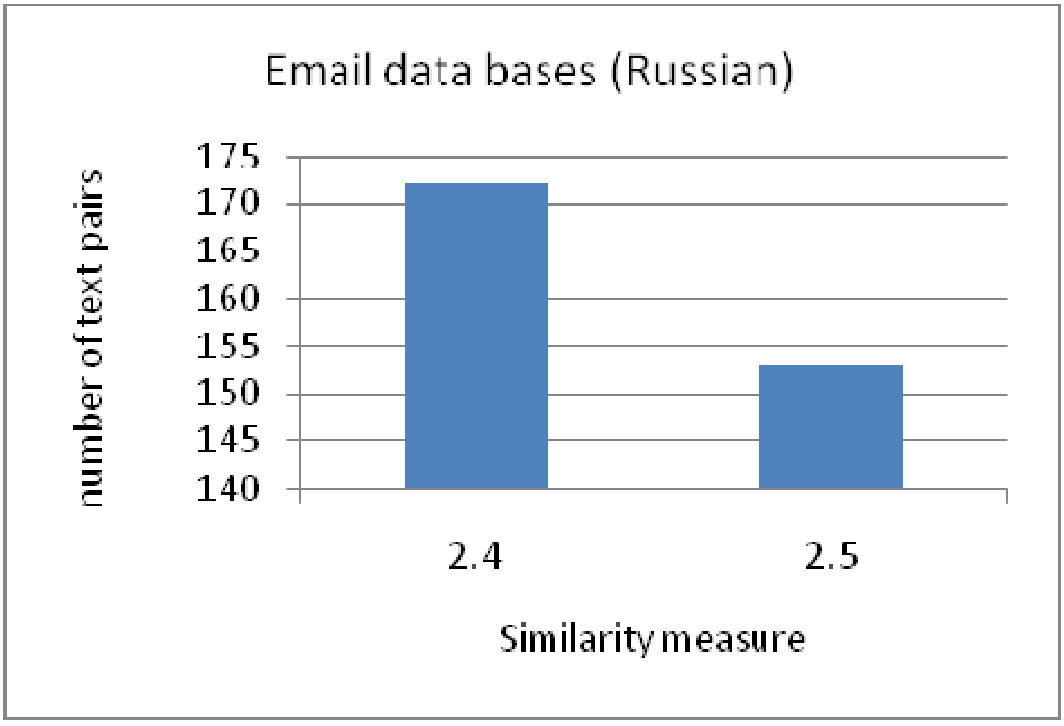

Fig. 4. Trigram similarity measure of "Email базы" mails

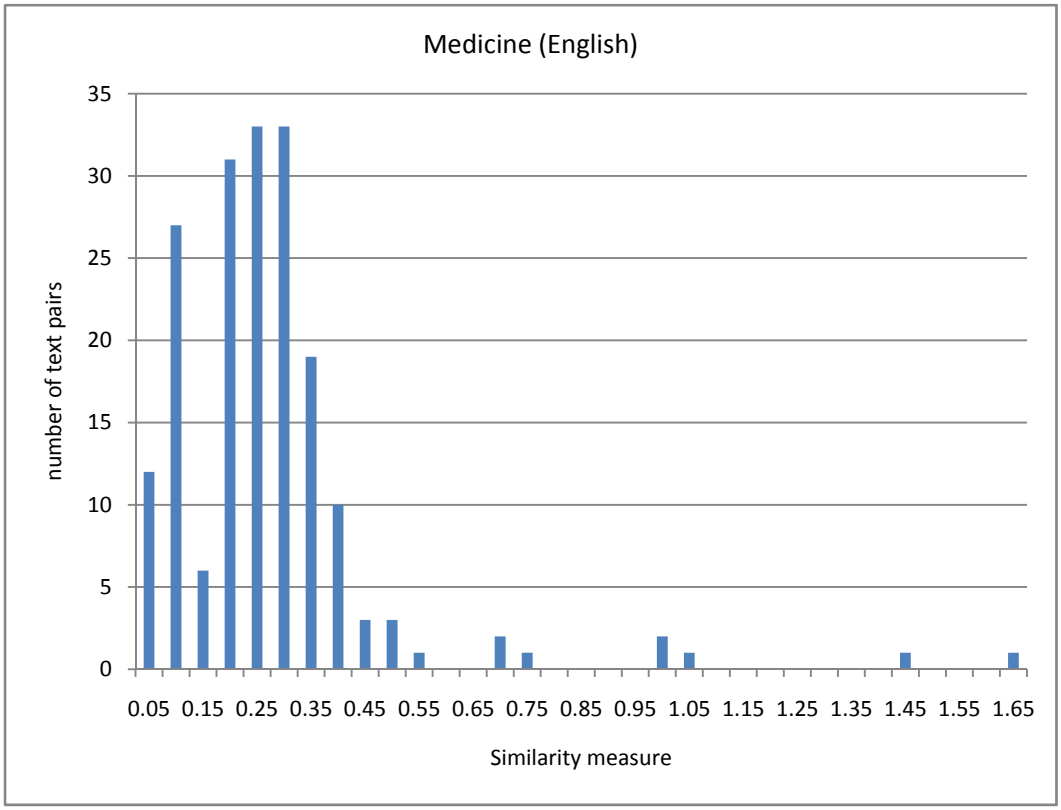

Fig. 5. Trigram similarity measure of "Medicine" mails

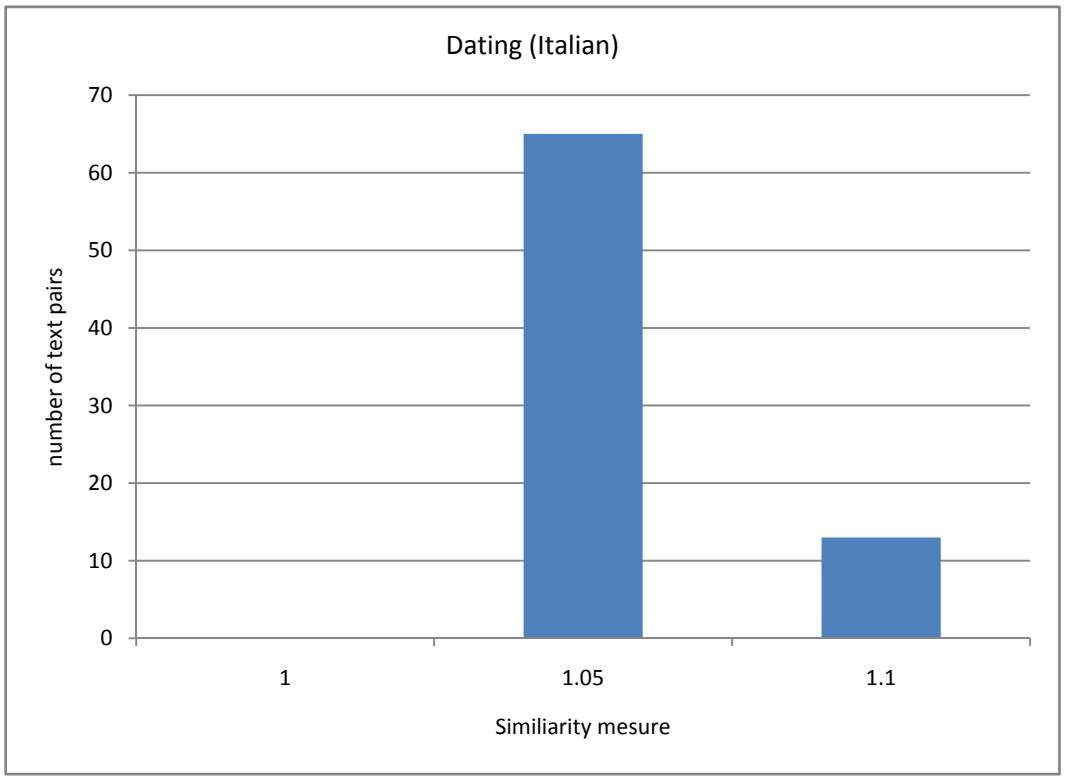

Fig. 6. Trigram similarity measure of "Dating" mails

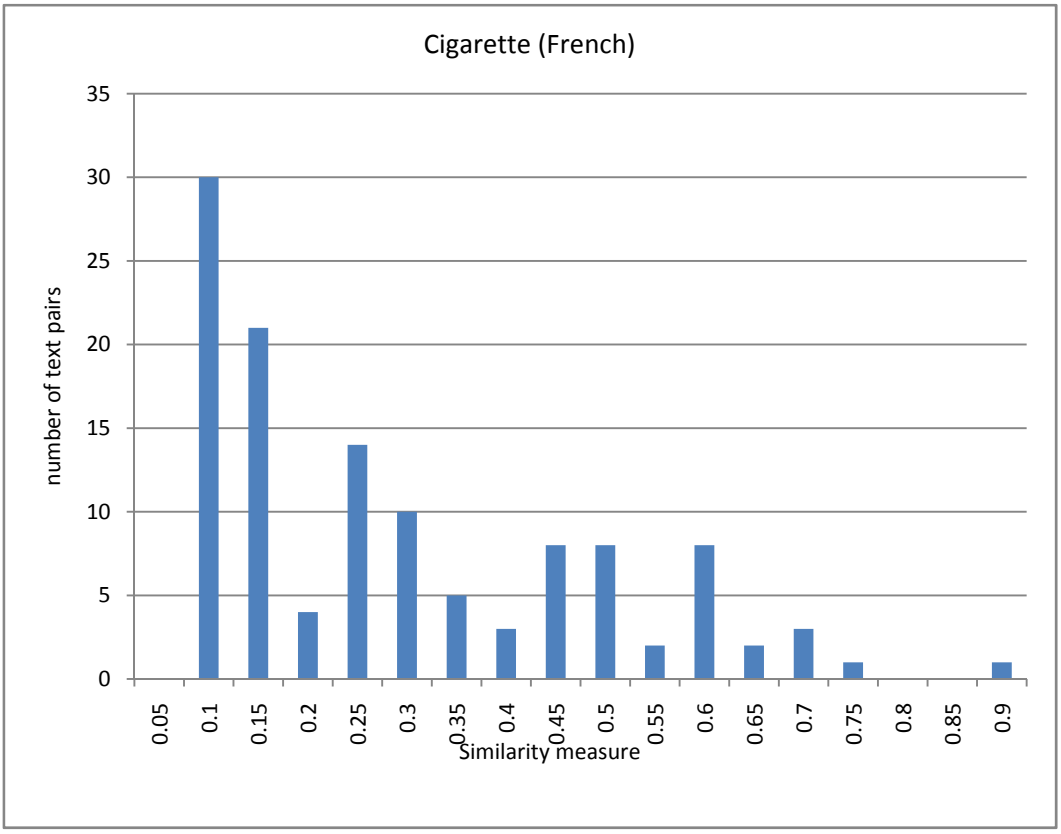

Fig. 7. Trigram similarity measure of "Cigarettes" mails

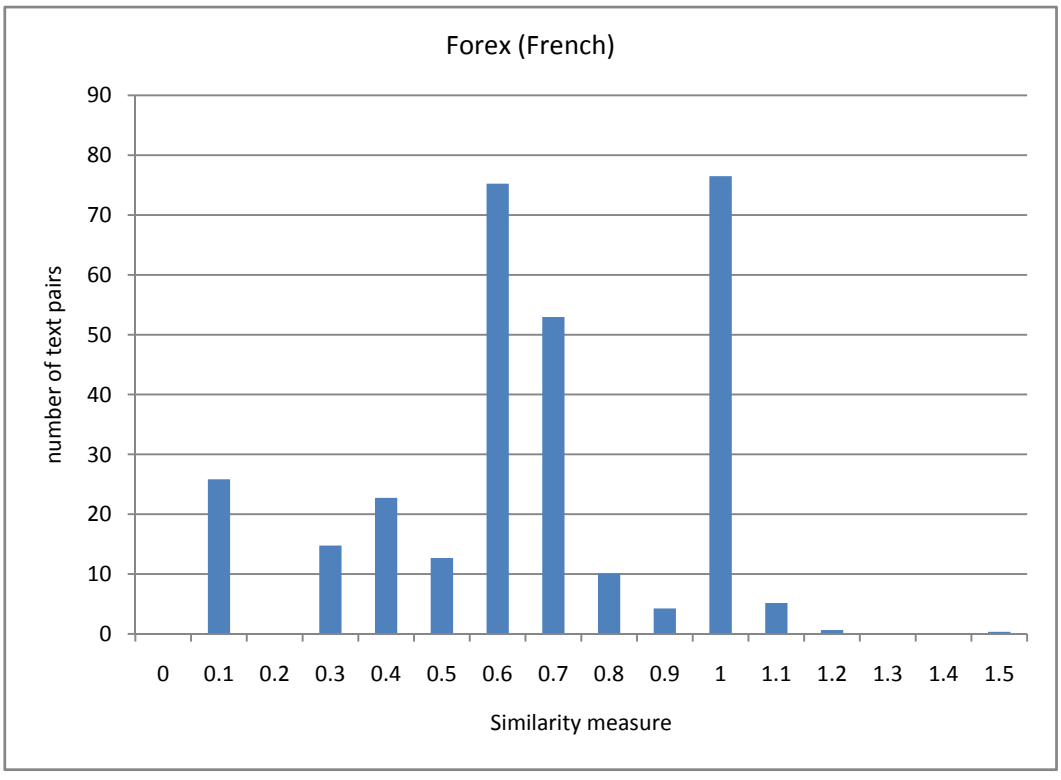

Fig. 8. Trigram similarity measure of "Forex" mails

It seems that trigram approach is not efficient because words may be rearranged. However, even in natural languages with flexible word order (e.g. Russian) there are syntagmatic regularities. Deviations from these regularities perform an emphatic function or make a text difficult to understand. Perception difficulties are not desired in spam since they reduce the response.

As a classifier a support vector machine (SVM) has been chosen. We have used STATISTICA 8.0. Quantitative features of Russian messages enable to identify transforming messages with high accuracy. SVM parameters are given in the Table 1. As we can see SVM detects transformers with extremely high accuracy. However, the results obtained by SVM may be checked by trigram method. It is possible to use other classifiers (e.g. neural networks are quite efficient).

Table 1. SVM parameters for the identification of Russian transforming messages

Sample size = 707 (Train), 236 (Test), 943 (Overall)
Support Vector machine results:
SVM type: Classification type 1 (capacity=10,000)
Kernel type: Radial Basis Function (gamma=0,007)
Number of support vectors = 118 (0 bounded)
Support vectors per class: 94 (0), 16 (1), 8 (2)
Class. accuracy (%) = 100,000(Train), 100,000(Test), 100,000(Overall)

Thus, there are three main steps of transforming messages detection:

1. Quantitative features retrieval;
2. Classification using SVM;
3. Trigram verification.

Phishing detection

According to definition given by Kaspersky Lab, "phishing is a type of Internet fraud that seeks to acquire a user's credentials by deception" [Kaspersky Lab, 2009]. It made up 0,57% in the first quarter of 2010 [Kaspersky Lab, 2010].

We have carried out a survey in order to establish what spam is considered to be, whether people think it is dangerous and how they protect themselves against it. 200 people have taken part in the survey. The respondents have been divided into several categories:

- According to profession/education: related to IT and non-related to IT
- According to age: 0-25, 25-40, more than 40
- According to sex.

Respondents should provide a spam definition and reveal its features. Moreover they have been asked about social engineering and identity theft.

59% of respondents consider spam as a kind of advertisement. 90% think that this phenomenon is not related only to email. 1% believes that any distributed advertisement (e.g. flyer) is spam. People have revealed the following spam indicators:

- 16% - links;
- 16% - unknown sender;
- 2% think that spam is the same thing as phishing;
- 45% consider spam as useless information;
- 11% define spam as nonsense messages.

The majority of respondents are not acquainted with concepts identity theft and social engineering. 50% do not consider social networks and IM to be dangerous.

A special phishing detecting software has been implemented. The algorithm is based on quantitative features and message structure. Moreover, we considered specific vocabulary related to phishing.

Here are the most frequent words occurring in phishing and other messages.

Table 2. Words occurred in phishing as well as in other messages

Word	Occurrence in phishing	Occurrence in notphishing
compte	55	592
paypal	53	3
carte	31	13
informations	29	52
images	24	224
cliquez	19	810
free	19	84
banque	16	38
visa	15	2
cher	13	71
client	13	23

Table 3. Words occurred only in phishing

Word	Occurrence in phishing
verified	10
activer	9
en_us	9
freebox	8
facturation	7
free.fr	7
desjardins	6
caisse	6
xxxx	6
accesd	5
suspendue	5

Table 4. Words not occurred in phishing

Word	Occurrence in notphishing
cliquez	810
compte	592
argent	583
dйmonstration	514
gagner	497
commencez	434
montres	384
йvиnements	372
prix	317
bonus	292
experts	273

securite	13	19
jour	12	106
service	11	63
lien	10	45
mettre	10	4
html	9	1594
passe	9	27
dessous	8	10
postale	8	2
ligne	7	84
information	7	9
confirmer	7	1
connexion	7	1
dernier	6	19
page	6	15
sites	6	8
limite	6	7
login	6	2

suspension	4
temporairement	4
suspendre	4
rappel	4
mesures	4
pixel	4
connecter	4
curitr	4
faveve	3
frauduleux	3
btn_org_arrow	3
contraints	3
bloquйe	3
populaire	3
protegez	3
kunstgeschichte	3
labanquepostale	3
inhabituelles	3
retablir	3

gratuitement	268
annonces	266
formation	263
assistance	256
grвce	251
intйgrale	245
arrкter	245
entraonez	244
toujours	232
images	224
ouvrir	222
atteindre	221
travailler	212
temps	210
sacs	208
bijourama	207
faire	199
forex	199
maintenant	184

As a classifier we also have used SVM algorithm built in STATISTICA 8.0.

Table 5. SVM parameters for phishing identification

Sample size = 994 (Train), 333 (Test), 1327 (Overall)
Support Vector machine results:
SVM type: Classification type 1 (capacity=10,000)
Kernel type: Radial Basis Function (gamma=0,009)
Number of support vectors = 57 (18 bounded)
Support vectors per class: 43 (0), 14 (1)
Class. accuracy (%) = 98,592(Train), 98,498(Test), 98,568(Overall)

Conclusion

Nowadays there are quite a lot of spam filters. Nevertheless, they are not efficient enough or they are very time and resource consuming. The majority of techniques are suitable only for email filtering. In contrast to them content methods may be applied to spam filtering in various message systems (IM, social networks etc.). The improvement of signature methods seems to be topical. The proposed techniques enable to identify transforming messages in a very efficient way. It is not extremely resource consuming as shingle approach and at the same time may be applied for various languages.

The performed survey allows drawing a conclusion that people underestimate threat related to IM, social networks and email. The majority of users are not familiar with the term "phishing". Proposed phishing detection technique is based on SVM. Quantitative characteristics, message structure and key words are used as features. Classification accuracy is above 98%. This approach may be improved by link analysis.

Bibliography

[Kaspersky Lab,2010] Kaspersky Lab, What spam is // Securelist, 2010, http://www.securelist.com/ru/encyclopedia/spam?chapter=151

[Namestnikova M.,2011] Namestnikova M. Spam v dekabre 2010 goda // Securelist, 2011, http://www.securelist.com/ru/analysis/208050676/Spam_v_dekabre_2010_goda

[Kaspersky Lab, 2010] Kaspersky Lab. Spam v pervom kvartale 2010 goda // Kaspersky Lab, 2010, http://www.kaspersky.ru/news?id=207733226

[Kaspersky Lab, 2009] Kaspersky Lab, Spam evolution // Securelist, 2009, http://www.securelist.com/ru/encyclopedia/spam?chapter=155

[Kaspersky Lab,2009] Kaspersky Lab, What is phishing? // Securelist, 2009, http://www.securelist.com/ru/encyclopedia/spam?chapter=155

[Yandex, 2010] Yandex, Nekotorye avtomaticheskie metody detektirovaniya spama dostupnye bolshim pochtovym sistemam // Yandex Company, 2010, http://company.yandex.ru/public/articles/antispam.xml

[Segalovich I., 2010] Segalovich I., Teyblum D., Dilevsky A. Principy i tekhnicheskye metody raboty s nezaprashivaemoy korrespondenciey // Yandex, 2010, http://download.yandex.ru/company/spamooborona-latest.pdf

[Kaspersky Lab,2009] Kaspersky Lab , Spamtest, 2009, http://www.kaspersky.ru/news?id=143937135

[Broder A, 2003] Broder A. On the resemblance and containment of documents // Digital Systems Research Center, 2003, http://ftp.digital.com/pub/Digital/SRC/publications/broder/positano-final-wpnums.pdf

[Manber U,1994] Manber U. Finding similar files in a large file system // USENIX Conference, 1994

[Chakrabarti S.,2003] Chakrabarti S. Mining the Web: Discovering Knowledge from Hypertext Data, 2003

[Coulthard M. 2004] Coulthard M. Author Identification, Idiolect and Linguistic Uniqueness. 2004

[Halteren H.,2004] Halteren H. Linguistic Profiling for Author Recognition and Verification// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004

[Sotnik S.,2006] Sotnik S. Identifikacya yazyka UNICODE teksta po N-grammam dlinoy do 4 vkluchitelno // Matematicheskoye modelirivanie, 2006, p. 111-114

[Cavnar W. B.,1994] Cavnar W. B., Trenkle J. M. N-Gram-Based Text Categorization // Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, 1994, стр. 161-175

[Mesheryakov R.,2005] Mesheryakov R., Vasukov N. Identifikacya avtora metodami iskustvennogo intellekta, 2005

[Fomenko V.,1983] Fomenko V., Fomenko T. Avtorsky invariant russkikh literaturnykh tekstov // Metody kachestvennogo analisa tekstov, 1983

[Rakhimova A.,2005] Rakhimova A. Lingvisticheskaya ekspertisa // Vestnik KASU, 2005

[Galyashina E.,2003] Galyashina E. Osnovy sudebnogo rechevedeniya, 2003

Author' Information

Liana Ermakova *– e-mail: liana87@mail.ru*

Major Fields of Scientific Research: text classification, filtering, information retrieval

COMPARATIVE ANALYSIS OF PHYLOGENIC ALGORITHMS [1]

Valery Solovyev, Renat Faskhutdinov

Abstract *The paper is dedicated to comparative analysis of phyogenetic algorithms used for linguistics tasks. At present there are a lot of phylogenetic algorithms; however, there is no unanimous opinion on which of them should be used. The paper suggests the model of language evolution trees and introduces a parameter to characterize the topology of trees. The comparison of the main algorithms is made on the trees of various topology. The paper displays that the UPGMA algorithm gives better results on the trees close to balanced ones. It provides the explanation for a number of contradictive results, described in published works.*

The problem of the input data choice and the relation between results and the number and type of parameters is under consideration. The results obtained are also ambiguous. Typological databases "Jazyki mira" and WALS as well as the method of computer modeling are used in the paper.

Keywords: l*anguage evolution, phylogenetic algorithms*

Introduction

In a number of papers [Nakhlen *et al.,* 2005-1, Nakhlen *et al.,* 2005-2, Cysouw and Comrie, 2009, Atkinson *et al.,* 2005, Donwey *et al.,* 2008, Wichmann and Saunders, 2007] attempts have been made to apply approaches developed in biology for reconstructing trees of species evolution to linguistic data. Recently compiled large databases like WALS [2005] and "Jazyki mira" [2011], ASJP [Müller *et al.,* 2010], which have introduced a great deal of new data for comparative research, hold the promise of producing new results in historic linguistics. The three databases are compared in [Polyakov *et al.,* 2009].

The phylogenetics suggests different algorithms for constructing evolutionary trees. Meanwhile the questions of better algorithm and better data are still open. The most popular phylogenetic algorithms in include UPGMA (Unweighted Pair Group Method with Arithmetic mean), NJ (Neighbour Joining), MP (Maximum Parsimony), and MrBayes.

The results described in published papers are contradictive. In the paper [Wichmann and Saunders, 2007] the NJ, MP, and Bayes algorithms were compared, and the last is considered to be the most suitable. In [Nakhlen *et al.,* 2005-1] the evolution of Indo-European family was studied and it was ascertained that NJ provides best

[1] The research was supported by Russian Foundation of Basic Research (grant № 10-06-00087-a.)

result. In fact the NJ algorithm has been recently used in linguistic researches. The belief in advantages of NJ algorithm is based on the paper [Saitou and Nei, 1987]. However, in [Donwey *et al.,* 2008] it was proved on the material of Sumba languages that UPGMA has better results. According to [Solovyev, 2011], algorithm NJ yields serious mistakes while applying the ASJP database. We compare these two algorithms as the most popular ones.

Another problem is data selection. The problem of choosing features for comparison is not trivial. In glottochronology the approach has been to only consider the most stable lexical items. A similar approach should be applied also to typological features. Attempts to define relative stabilities for WALS features are presented in [Wichmann and Kamholz, 2008] and, with improved methods, in [Wichmann and Holman, 2009].

The paper considers dependence on a number of used features and their type (i.e. what part of grammar they belong to). Besides, dependence of the results on stability of features is analyzed with the use of the "Languages of the World" database.

Comparison of algorithms

Careful analysis of the argumentation given in paper [Saitou and Nei, 1987] shows that NJ provides better results on the trees of a certain topology (= structure). As a matter of fact the authors of the paper tested only two very specific topologies of trees. Besides, the research in [Saitou and Nei, 1987] was initially oriented to the studies of biological evolution, but not a language one. The trees of a language family level are not usually like these ones. That is why the task of systematic comparison of the algorithms on the trees of different configuration is of vital importance as well as the constructing the realistic model of language evolution trees.

We analyze different cases of using the algorithms NJ and UPGMA, showing that UPGMA often gives better results than NJ in the certain cases. The influence of the tree topology on the result is being studied. Comparison of trees from papers [Nakhlen *et al.,* 2005-1] and [Donwey *et al.,* 2008] let us hypothesize that if a reconstructed tree is close to the balanced one (all branches have the same number of edges) UPGMA can be more accurate than NJ.

First of all we propose the model of language evolution trees. We studied the question of edges length variations in the real trees of language evolution. One of the most completely described trees is the evolution tree of the Turkic family, given in paper [Sravnitel'no-istoricheskaja, 2002]. The lengths of all edges in the tree (there are 77 of them) have been calculated and located in the order of increasing. The results are represented in Diagram 1.

It turned out that there are several super long edges. The longest, which is of 2130 years, corresponds to the initial separation of the Chuvash language from proto-Turkic language. The next longest edges (1330 and 1270 years) demonstrate separating the Yakut language from the Siberian branch and the Salar language from the Oguz branch. There is one abnormally short edge of 30 years that is the edge in evolution tree of Kypchat languages. The lengths of the majority of edges excluding the shortest and the ten longest edges can be strictly

put on the direct line. The fact that the lengths of the majority of edges except some of them can be put on the direct line means that the edge lengths can be considered as a random value with an even distribution.

The lengths vary from 90 to 650 years. Thus, the average meaning of an edge length is 370 years. The declination is ± 280 years that equals 75% average length. Similar results are obtained for other language families. This data is a basis for the algorithms of random tree generation below.

We conducted an experiment with generation of random binary trees of arbitrary topology to check the hypothesis. The trees were generated with a given number of leaves and the length of each edge was determined as a random number on a given interval. Then, matrixes of distances between leaves were made for every generated tree T. After that, trees T-UPGMA and T-NJ were determined by methods UPGMA and NJ.

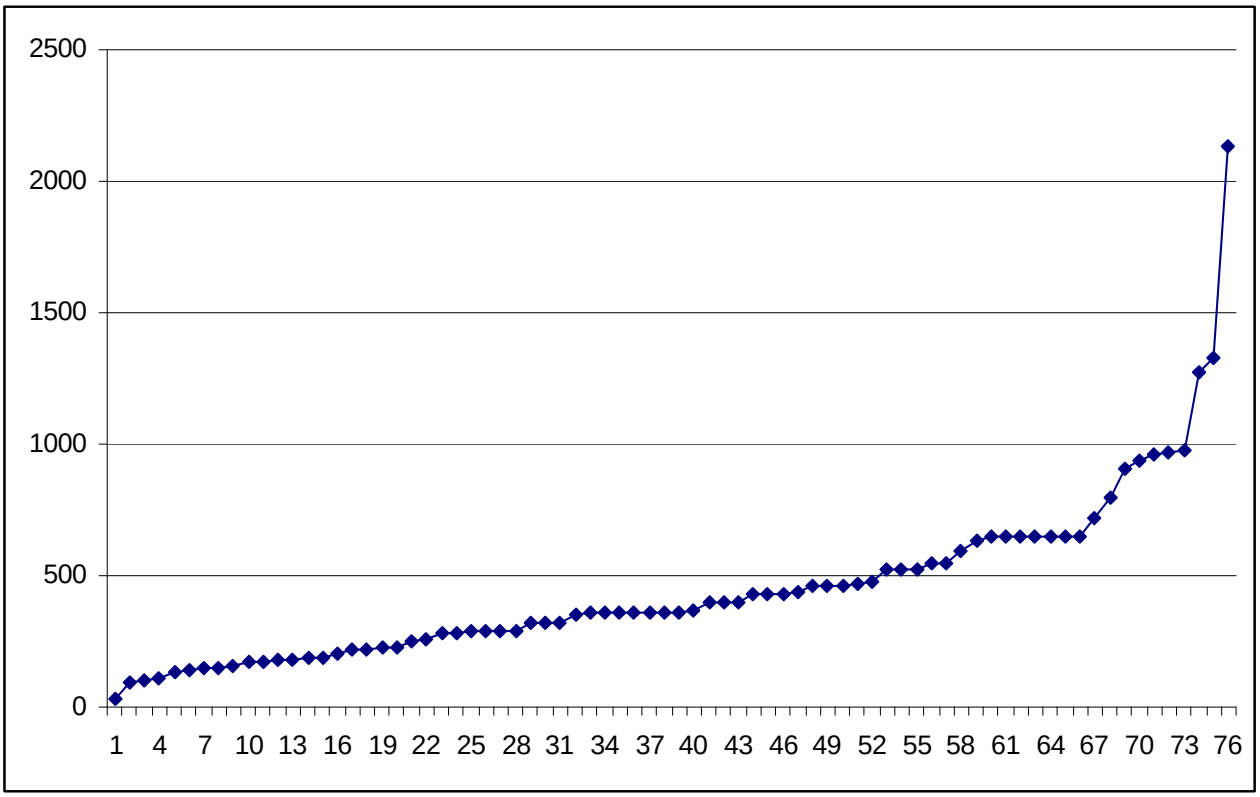

Diagram 1. Lengths of edges of the evolution tree of Turkic languages

In order to assess how big the difference between two trees is we used the Robinson-Foulds distance [Pattengale *et al.,* 2007] between them, which stands for the number of elementary transformations needed for conversion of one tree into another.

The measure of branching, as the sum of levels of inner nodes, is introduced to characterize numerically the degree of closeness of the tree to the balanced tree. In this case the root level equals 0 and the level of an ancestor is greater than the level of a descendant by 1. It is obvious that the closer a tree to the balanced tree, the smaller the measure of its branching.

To be more precise one can describe the whole algorithm as the following:

1. A random binary tree T with a given number of leaves r is generated, and all the edges are of equal length 1.

2. The measure of branching of tree T is calculated.

3. In the generated tree the length of each edge is changed by a random number from interval [– p, + p], where p = 0.75.

4. The matrix of distances between leaves is constructed by the data.

5. Trees T-UPGMA and T-NJ are constructed by the distance matrix by methods UPGMA and NJ.

6. The Robinson-Foulds distance between the obtained trees and tree T is calculated.

We made calculations for two cases, when number of leaves is equal to 15 and 50. 1000 random trees have been generated and the results have been averaged. The branching measure for the trees with 15 leaves is from 31 to 105 and for the trees with generated random sample it was from 33 to 58. It is convenient to divide all the trees by the measure of their branching into several groups in order to analyze the data obtained. We chose four groups approximately equal by the number of trees with the following values of measure: 31-36, 37-40, 41-45, 46-105. For each group we calculated the averaged Robinson-Foulds distances, given in the Table 1.

Table 1. Averaged distances, r = 15 leaves

Measure of branching	UPGMA	NJ
31 - 36	4,31	5,04
37 - 40	6,41	5,72
41 - 45	8,11	6,42
46 - 105	9,04	7,43

It is clear that the efficiency of the algorithms depends on the topology of trees. For trees with a small measure of branching, which are close to a balanced one, better results are provided by UPGMA algorithm. The similar result is obtained for r = 50.

Thus, it has been proved that NJ algorithm is not undoubtedly the best one. Both real examples and modeling by generation method of random trees shows that UPGMA is preferable in a number of cases.

Data selection

We begin handling the problem of selection of features from the WALS-based investigation of a number and type of features for algorithms UPGMA, NJ, and MP.

We used the following six pairs Americas languages from six different families (also considered in [Wichmann and Saunders, 2007]):

1. Athapaskan: Slave, Navajo
2. Uto-Aztecan: Yaqui, Comanche
3. Chibchan: Ika, Rama,
4. Aymaran: Aymara, Jaqaru
5. Otomanguean: Chalcatongo Mixtec, Lealao Chinantec
6. Carib: Hixkaryana, Carib.

We tried to reveal the dependence on the number and the type of features using this language set. Having used all the set of WALS structural features (142 feature) as well as 60 randomly chosen features (i.e. a bit less than a half of them all) we obtained the following results. Random features were chosen three times, and the average data are represented. Following [Wichmann and Saunders, 2007], we use also the 17 best features.

Table 2. Dependence on a number of features

Algorithms	17 Features	142 Features	60 Features
UPGMA	5	4	3,7
NJ	3	3	4,3
MP	4	4	4,3

A complete set of features gives a slightly inferior result, but is comparable with the set of highly-informative features, selected in [Wichmann and Saunders, 2007]. A reduced number of features (up to 60) leads to sharp change for the worse of results for UPGMA. At the same time the results for NJ and MP algorithms improved. It means that the quality of the algorithm results strongly depend on a number of features that needs further investigation. The algorithms having been analyzed are strictly divided into two groups: UPGMA and NJ, MP. The latter group works better with an average number of features.

Table 3. Dependence on a type of features

	Phonetic features	Morphological features	Syntax features
UPGMA	3	1	4
NJ	2	4	6
MP	2	4	4

The next experiment was aimed at explanation of the contribution made to general classification by separate aspects of language such as phonetics, morphology and syntax. The data are given in Table 7. Phonetic features are the features 1-19 WALS, morphological features – 20-56, syntax features – 57-128 (other WALS features are not grammatical).

It was unexpected to some extent that good results were obtained for a set of syntax features. The great expectations were connected with morphological characteristics, since they are presumably less borrowable. That is why one could expect that they would be more useful for explanation of genetic relations. On the other hand, many syntax properties change very slowly. J. Nichols [2007] suggested using some of them for establishing genetic relations.

Let us consider the ways how feature stability influents the result. General information on grammatical features' stability is available from [Wichmann & Holman [14]. We use the database "Jazyki mira". 503 most informative features (that are found at least in 25 languages but no more than in 300 languages) were selected. 4 measures for feature stability were under consideration: [Maslova, 2004], [Nichols, 1995], Wichmann & Holman [14], [Solovyev and Faskhutdinov, 2009].

For every measure the features were divided into four approximately equal by number of features groups, from the maximum (group 1) to the minimum (group 4) degree of stability. For every feature group we constructed evolution trees by NJ algorithm. The Robinson-Foulds distances were calculated between consensus tree (for languages from "Jazyki mira") and the trees constructed by NJ for all stability groups. The results are in the table 4.

Table 4. Robinson-Foulds distances for different measures and stability groups

Stability measure/Group Number	Group 1	Group 2	Group 3	Group 4
Maslova's measure	52	54	50	44
Nichols's measure	48	46	54	46
Wichmann's measure	50	52	54	44
Solovyev's measure	50	50	52	40

Best trees are constructed in the fourth group (with the lowest degree of stability) for all stability measures. The obtained result can be explained by the fact that the features from the first three groups are less informative.

Conclusion

More and more wide application of philogenetic algorithms in linguistic studies calls for consideration of justification of choice of both algorithms and data. Despite the existence of a number of methodological publications, first of all, the abovementioned [Wichmann and Saunders, 2007], many open questions remain.

The paper suggests the model of a language evolution trees and introduces the measure of trees' balance. In a number of cases, namely, for almost balanced trees, based on the model comparison of NJ and UPGMA algorithms proved higher efficiency of UPGMA. This provides theoretical explanation for a number of previously published results.

Consideration of several ways of selection of features proved the expediency of an increased attention to syntactic features, which are moderately persistent. Far from being exhaustive, the conducted research hints at promising venues of future undertakings.

Bibliography

[Atkinson *et al.,* 2005] Atkinson Q., Nicholls G., Welch D., Gray R.: From words to dates: water into wine mathemagic or phylogenetic inference? Trans. of the Philological Society. V.103:2, 2005. p.193-219.

[Donwey *et al.,* 2008] Donwey S., Halmark B., Cox M., Norquest P., Lansing S. Computational Feature-Sensitive Reconstruction of Language Relationships: Developing the ALINE Distance for Comparative Historical Linguistic Reconstruction. Journal of Quantitative Linguistics. V.15, N4, 2008, pp. 340-369.

[Maslova, 2004] Maslova E. Dinamika tipologicheskih raspredelenij i stabil'nost' jazykovyh tipov. Voprosy jazykoznanija. № 5. 2004. C. 3–16. (In Russian).

[Müller *et al.,* 2010] Müller, André, Søren Wichmann, Viveka Velupillai, Cecil H. Brown, Pamela Brown, Sebastian Sauppe, Eric W. Holman, Dik Bakker, Johann-Mattis List, Dmitri Egorov, Oleg Belyaev, Robert Mailhammer, Matthias Urban, Helen Geyer, and Anthony Grant. 2010. ASJP World Language Tree of Lexical Similarity: Version 3 (July 2010). http://email.eva.mpg.de/~wichmann/language_tree.htm.

[Nakhlen *et al.,* 2005-1] Nakhlen L., Ringe D., Warnow T.: Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. Language. v. 81, 2005. p. 382-420.

[Nakhlen *et al.,* 2005-2] L. Nakhleh, T. Warnow, D. Ringe, and S.N. Evans, A Comparison of Phylogenetic Reconstruction Methods on an IE Dataset. The Transactions of the Philological Society, 3(2): 171-192, 2005.

[Nichols, 1995] Nichols J. Diachronically stable structural features. Andersen, Henning (ed.), Historical Linguistics. 1993. Selected Papers from the 11th International Conference on Historical Linguistics. Los Angeles 16–20 August 1993. Amsterdam/Philadelphia: John Benjamins Publishing Company. 1995. P. 337–355.

[Nichols, 2007] Nichols J. Typology in the service of classification. http://aalc07.psu.edu/ papers/jn_typol_class3.pdf. Stanford, 2007.

[Pattengale *et al.,* 2007] Pattengale N., Gottlieb E., Moret B. Efficiently Computing the Robinson-Foulds Metric. - Journal of Computational Biology. 2007, 14(6): 724-735.

[Polyakov *et al.,* 2009] Polyakov V., Solovyev V., Wichmann S., Belyaev O. Using WALS and Jazyki Mira. Linguistic typology. 2009. v. 13, № 1.

[Saitou and Nei, 1987] Saitou N., Nei M. The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. Mol. Biol. Evol. V.4, N4, 1987. pp 406-425.

[Solovyev, 2011] Solovyev V. The Problem of Interpretation of Phylogenetic ASJP Trees. ASJP-project, 2011, http://email.eva.mpg.de/~wichmann/papers.htm

[Solovyev and Faskhutdinov, 2009] Solovyev V., Faskhutdinov R. Metodika otcenki stabil'nosti grammaticheskih svojstv. Izvestija RAN. V. 68., № 4., 2009. (In Russian).

[Sravnitel'no-istoricheskaja, 2002] Sravnitel'no-istoricheskaja grammatika tjurkskih jazykov. Red. E.R.Tenischev. Moscow: Nauka. 2002. (In Russian).

[Cysouw and Comrie, 2009] M. Cysouw, B. Comrie. How varied typologically are the languages of Africa? In: Rudie Botha & Chris Knight. (eds.) The Cradle of Language. Volume 2. Oxford, 2009.

[Jazyki mira, 2011] The Database "Jazyki mira". 2011, http://www.dblang.ru.

[WALS 2005] The World Atlas of Language Structures. Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie (eds.). Oxford: Oxford University Press, 2005. 695 p.

[Wichmann and Holman, 2009] Wichmann S., Holman E. Temporal stability of linguistic typological features. Lincom Europa: Muenchen. 2009.

[Wichmann and Kamholz, 2008] Wichmann S., Kamholz D. A stability metric for typological features. STUF – Language Typology and Universals. 2008. v. 61. p.251-262.

[Wichmann and Saunders, 2007] Wichmann S., Saunders A. How to use typological database in historical linguistic research. Diachronica. 2007. v. 24, №2.

Author's information

Valery Solovyev *– Professor, Department of Computer Science, Kazan Federal University, Kremlevskaja, 18, 420008 Kazan, Russia; e-mail: maki.solovyev@mail.ru*

Major Fields of Scientific Research: Cognitive linguistics, Language evolution, Quantitative linguistics

Renat Faskhutdinov *– Junior Researcher, Department of Computer Science, Kazan Federal University, Kremlevskaja, 18, 420008 Kazan, Russia; e-mail: jvenal@mail.ru*

Major Fields of Scientific Research: Quantitative linguistics, Linguistic databases

ANALYZING THE LOCALIZATION OF LANGUAGE FEATURES WITH COMPLEX SYSTEMS TOOLS AND PREDICTING LANGUAGE VITALITY

Samuel F. Omlin

***Abstract**: Half of the world's languages are in danger of disappearing before the century ends. Efficient protection of these languages is difficult as their fate depends on multiple factors. The role played by the geographic situation of a language in its survival is still unclear. The following quantitative study focused on the relation between the 'vitality' of a minority language and the linguistic structure of the neighboring languages. A large sample of languages in Eurasia was considered. The languages were described based on a complex system of typological features. The spatial distribution of the language features in the sample area was measured by quantifying deviations from purely random configurations. Interactions between the linguistic features were revealed. The obtained interaction network permitted to define a location "quality" index for a language localization. This index was put in relation to corresponding vitality estimations from Unesco. A significant relation could be established between these two variables. The degree of endangerment of the minority languages studied seems effectively related to the linguistic structure of their neighboring languages. Beyond the particular context of endangered languages, the proposed approach constitutes a promising tool to gain more knowledge about the mechanisms that control the geographical distribution of linguistic features.*

***Keywords**: Language competition, Complex systems, Interactions, Spatial distribution, Typological language features.*

***ACM Classification Keywords**: I.m Miscellaneous; J.5 Arts and Humanities – Linguistics; H.2.8 Database Applications – Data mining, Scientific databases, Spatial databases and GIS.*

Introduction

Numerous languages of the world are endangered: Unesco [2010] estimates that half of the about 6700 languages spoken today will have disappeared before the century ends, if no significant measures are taken. However, efficient revitalization of endangered languages is a challenging task as the fate of a language depends on a myriad complex factors.

In the business world, numerous factors determine the success or failure of an enterprise. Yet, a well-known business doctrine states that the location alone determines, in most cases, whether or not a store may survive the

competition. This study attempts to determine if this doctrine can be applied to the survival of languages, i.e. if the localization of a language is a determinant factor in predicting the success of a language in competition with its neighboring languages. In fact, the following brief overview of recent models of language competition and some comparison to linguistic literature show that the role played by the geographic situation of a language in its ultimate survival, and in particular the role of the linguistic structure of the languages neighboring it, is still unclear.

Abrams and Strogatz [2003] proposed a simple model that describes how two neighboring languages are "competing for speakers". Their article received the attention of numerous researchers because "its fit to the empirical data was exceptional" [Wang and Minett, 2005, p. 265]. The proposed model predicted the death of the weaker language in every case. Many other language competition models followed. While some produce similar results [e.g. Castelló, Eguíluz, and Miguel, 2006, Castelló et al., 2007], others lead to the conclusion that under certain circumstances, minority languages can live in stable coexistence with stronger ones. Models allowing stable coexistence notably focus on historical or geographical factors [Patriarca and Leppänen, 2004, Patriarca and Heinsalu, 2009], include the pride of speakers to their linguistic identity [Schulze and Stauffer, 2006], or take into account linguistic similarity between two or more languages "in competition" [Mira and Paredes, 2005, Teşileanu and Meyer-Ortmanns, 2006]. The following comparison of these results with linguistic literature focuses on the role of the linguistic structure of languages in competition.

Mira and Paredes [2005, p. 1] who studied the coexistence of Galician (in northern Spain) with the dominant Castilian, concluded that two languages can coexist if they are "similar enough". Teşileanu and Meyer-Ortmanns [2006] drew the same conclusion about two or even three neighboring languages. These results are to some extent in agreement with Wardhaugh [1987, p. 17], who notes that in certain situations "there may be little pressure from one language on the other or others". Other linguists affirm that a high linguistic similarity between a minority language and a stronger one can "retard" its assimilation, notably based on the case of the Galician [Monteagudo and Santamarina, 1993], as well as on an a example in the Netherlands [Palmer, 1997] and in Italy [Posner and Rogers, 1993]. However, Posner and Rogers [1993, p. 55] consider this case as an exception and state: "The greater the linguistic distance between languages the less likely is language shift to occur" [similarly Mackey, 2001].

A group of experts from Unesco estimated the degree of endangerment of the living languages in the *Atlas of the World's Languages in Danger* [Moseley, 2009], referred to as *language vitality*, with nine criteria. However, in none of them was geography directly implied. This stresses the importance of clarifying the influence of geographic factors in the context of language endangerment.

The following quantitative study focused on the relation between the vitality of a minority language and the linguistic structure of the languages neighboring it. For this purpose, a mathematical method, having its origins in the economical sciences and identifying optimal localizations to implement commercial stores with empirical success [Jensen, 2006, 2009], was adapted. In fact, Jensen had developed an approach to study the distribution of commercial activities in a city – a network on a heterogeneous geographic space. Similarly, the world is home to a network of languages, or more precisely, of linguistic features. The adapted approach was applied to a big sample of Eurasian languages, the linguistic structure of which was quantified based on a complex system of typological features. The present study is the first, known to the author, to integrate realistic linguistic features in order to describe languages in competition. Teşileanu and Meyer-Ortmanns [2006] had stressed the importance of such an approach. It is also the first to consider a large-sized language network in this context.

After this introduction, the paper is structured as follows: the next section presents briefly the studied language sample and its modeling; section 3 summarizes Jensen's approach [2006, 2009]; section 4 explains an approach to measure the spatial distribution of linguistic features; section 5 shows how the vitality of minority languages can be predicted based on these measurements; section 6 presents the most important results; the last section concludes the study and lists future work to do.

2 Sample and Modeling

A sample of 105 living languages in Eurasia was considered. For this study, the definition and usage of the term *language* of Ethnologue, 16th edition [Lewis, 2009] was used. This allowed referring to a language by a unique key, the code ISO 639-3. The *World Language Mapping System*[1] [Global Mapping International and SIL International, 2010; abbreviated as *WLMS]* permitted to describe the geographical area where the considered languages are spoken and to separate these languages into 186 linguistic communities with *independent vitality*. The term *linguistic community* was defined as the ensemble of speakers of a language in the same country. In fact, speakers of the same language but living in different countries do not have the same political and social environment, such that the speakers of a country can be considered as an independent linguistic community with an independent fate. Such a linguistic community could be identified by the unique code ID-ISO-A2 available in WLMS. This key is a concatenation of the code ISO 639-3 to identify a language and the country code ISO A2. The linguistic structure of a language was quantified based on the database *Jaziky mira* [Academia and Indrik, 1993-2010; abbreviated as *JM*], providing a complex hierarchical system of typological features. Every

[1] Version 16 (elaborated based on Ethnologue, 16th edition)

considered language was characterized by a certain number of *binary typological features*[1]. For the considered linguistic communities, the description of their language in JM could be univocally attributed via the code ISO 639-3. Additionally, a number of speakers (provided by WLMS) was associated and where possible as well a *vitality grade*[2] from the *Atlas of the World's Languages in Danger* [Moseley, 2009; further referred to as *Unesco's atlas*]. The attribution of vitality grades to linguistic communities was a complex and fastidious task as Unesco does not use a standardized key in order to refer to them. In addition, it eventuated that about two thirds of the examined *central points*[3] from Unesco could not be linked to an area from WLMS (nor to a set of areas). The first step consisted in extracting the 186 central points from Unesco's atlas that refer to one single language (i.e. having associated one and only one code ISO 639-3), which is described in JM (the other languages were not of interest for this study). Then, for 43 linguistic communities a clear link to a vitality grade could be established on the basis of some geographic criteria (mainly: global geographic situation, distance of the Unesco's central point to the area and central point of WLMS) as well as on a comparison of the metadata of the two databases WLMS and Unesco's atlas (name(s) of the attributed language, number of speakers, textual description of localization, etc.). 31 of these communities were in the chosen *sample region* (explained in the following). The modeling described above is summarized in the (spatial) UML schema in figure 1.

The three most important objectives when defining the sample were the following: the sample has to be in a *continuous geographic region*; it should represent an *inventory* of linguistic communities of the chosen region *as complete as possible*, especially of the communities with many speakers (evidently, a community could only be included in the sample if it could be described by JM); nevertheless, the sample should be of a *considerable size*, in order to a priori allow gaining statistically significant results. As the two first objectives were concurrent to the last one, an optimal compromise was aimed at. The retained sample region consists of nearly entire Europe, a major part of Asian Russia and a few adjacent regions, in particular the region of the Caucasus Mountains – a "hotspot" of endangered languages – together with its close surroundings.

---

[1] An example of such a binary topological feature is the following: "The word order in the simple phrase is subject, followed by verb and then by object." For this feature, English obtains the value '1' or 'True', Turkish the value '0' or 'False'. This example does not exist exactly like this in JM. It was invented for allowing an easy explanation of the structure of this database. Only the leaves of the hierarchical tree of JM were considered for characterizing the structure of the considered languages.

[2] Unesco attributed to every inventoried linguistic community one of the following ordinal vitality grades: "extinct", "critically endangered", "severely endangered", "definitely endangered" or "vulnerable". When a language is constituted of several independent communities, an *overall vitality* was attributed to the language.

[3] Represents the centre of the area where the speakers live or the coordinates of the largest city or village.

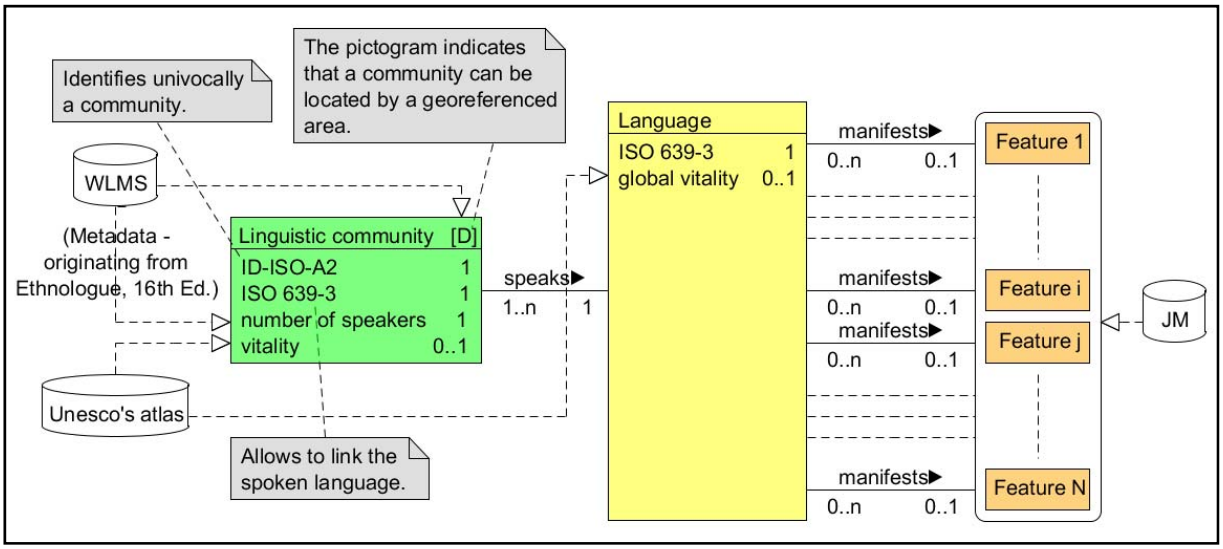

Figure 1: UML schema summarizing the modeling.

3 Analyzing a network on a heterogeneous geographic space

Jensen [2006, 2009] developed an approach to measure the spatial organization of commercial activities in a city. Concretely, he conceived an *M index* to quantify the geographic aggregation and dispersion tendencies of categories of stores. Jensen [2006, p. 1] explained:

The definition of M_{AB} at a given distance r is straightforward: draw a disk of radius r around each store (s) of category A, count the total number of stores (n(s)), the number of B stores ($n_B(s)$), and compare the ratio $n_B(s)/n(s)$ to the average ratio N_B/N, where N refers to the total number of stores in town. If this ratio, averaged over all A stores, is larger than 1, this means that A "attracts" B, otherwise there is repulsion between these two activities. I have chosen r = 100m as this represents a typical distance a customer accepts to walk to visit different stores.[1]

[1] In this quotation Jensen's notation of the variables has been slightly modified (the variable s has been inserted for referring to a store and "*n(S)*", "$n_B(S)$" and "***N***" have replaced "n_{tot}", "n_B" and "N_{tot}" of Jensen's notation). This was done to allow using the presented notation in the explicit formulas that follow.

The ratio $n_B(s)/n(s)$ can be understood as the *concentration* of activity B in the *neighborhood* of A and N_B/N as the *overall concentration* of activity B in the entire city [Prof. Pablo Jensen, oral communication, 2010].

After having presented the basic idea of the M index, Jensen [2009] presented two explicit formulas for the M index, *M intra* and *M inter* (M_{AA} and M_{AB}):

$$M_{AA} \equiv \frac{\frac{1}{N_A}\sum_{s\in S_A}\frac{n_A(s)}{n(s)}}{\frac{N_A-1}{N-1}} \tag{1}$$

$$M_{AB} \equiv \frac{\frac{1}{N_A}\sum_{s\in S_A}\frac{n_B(s)}{n(s)-n_A(s)}}{\frac{N_B}{N-N_A}} \tag{2}$$

where S_A refers to all A stores in the entire city, $n_A(s)$ represents the number of A stores in the neighborhood of store s and N_A stands for the number of A stores in the entire city. The elements in the formula for M inter that were not explained above constitute some corrections that only are of importance in extreme cases. The basic idea of the index remains the same. The basic idea behind M intra is analog to the one of M inter. It can also be interpreted similarly: "if the observed value of the intra coefficient is greater than 1, we may deduce that A stores tend to aggregate, whereas lower values indicate a dispersion tendency" [Jensen, 2009, p. 13]. The interpretations of M intra and M inter are based on the fact that under *pure randomness hypothesis* $E[M_{AA}]$ and $E[M_{AB}]$ equal 1 for all $r>0$.

Based on this M index, Jensen defined a location "quality" index ($Q_A(x,y)$) for a commercial activity A at a point (x,y) as:

$$Q_A(x,y) \equiv \sum_{B\in Act} a_{AB} n_B(x,y) \tag{3}$$

where Act is the set of all considered commercial activities, $n_B(x,y)$ represents the number of neighbor stores of category B around (x,y), $a_{AB}\equiv\log(M_{AB})$ for $A\neq B$ and $a_{AB}\equiv\log(M_{AA})$ for $A=B$. Jensen [2009, p. 18]

explained on the example of bakeries: "The basic idea is that a location that gathers many activities that are 'friends' (i.e. that attract bakeries) and few 'enemies', might well be a good location for a new bakery." The *location quality* for an existing store can be computed by removing it from town and calculating Q at its location [Jensen, 2006, 2009].

4 Measuring the spatial distribution of linguistic features

Jensen's M index [2009] was adapted in order to quantify tendencies of typological language features to aggregate or disperse. The *neighborhood of a linguistic community* was defined as the set of communities overlapping its area enlarged by a buffer of size r (the area of a community can be contained entirely by another one…). For easiest understanding the main idea of the index is explained in Jensen's words: for each community (c) manifesting the feature A, sum the number of speakers of all communities in its neighborhood ($n(c)$), sum the number of speakers of the communities manifesting the feature B in its neighborhood ($n_B(c)$), and compare the concentration $n_B(c)/n(c)$ to the overall concentration N_B/N, where N refers to the sum of the number of speakers of all communities in the entire sample region. If this ratio, averaged over all communities manifesting feature A (where the average is weighted by their number of speakers), is larger than 1, this means that A "attracts" B, otherwise there is repulsion between these two features. The buffer size r has been chosen 1 degree ($\approx$ 110 km). This size lies somewhere in between the maximal commute distance to work undertaken on daily basis and the distance that can be reached by car, train or autobus on a day trip. This choice may appear somewhat arbitrary, but the model seems quite robust to the variations of this parameter [Prof. Jensen, oral communication, 2010].

The spatial distribution of the language features was measured by counting speakers manifesting certain features rather then simply counting communities manifesting them. In fact, this allowed measuring the distribution with higher precision. It seemed important since the number of speakers of the considered communities vary very strongly.

The explicit formulas M intra and M inter were adapted:

$$M_{AA} \equiv \frac{1}{N_A} \sum_{c \in C_A} n_{self}(c) \frac{\frac{n_A(c)}{n(c)}}{\frac{N_A - n_{self}(c)}{N - n_{self}(c)}} \tag{4}$$

$$M_{AB} \equiv \frac{\frac{1}{N_A} \sum_{c \in C_A} n_{self}(c) \frac{n_B(c)}{n(c) - n_A(c)}}{\frac{N_B}{N - N_A}} \tag{5}$$

where C_A refers to all communities manifesting feature A in the entire sample, $n_A(c)$ represents the sum of the number of speakers of all communities manifesting feature A in the neighborhood of the community c, $n_{self}(c)$ equates to the number of speakers of community c itself and N_A represents the sum of the number of speakers of all communities manifesting feature A in the entire sample region ($N_A \equiv \sum_{c \in C_A} n_{self}(c)$). The elements in the formula for M_{AB} that were not explained above constitute the analogue corrections to the ones Jensen (2009) made. The differences of the adapted formula for M_{AA} to Jensen's original one were necessary to achieve the analogue corrections[1]. Also in these adapted formulas the corrections are only of importance in extreme cases and the basic idea of the index remains the same. Likewise, the main idea behind M intra is analog to the one of M inter and can again be interpreted similarly: if the observed value is superior to 1, it means that communities manifesting the feature A tend to aggregate, whereas inferior values indicate a dispersion tendency. As for Jensen's formulas, under pure randomness hypothesis, $E[M_{AA}]$ and $E[M_{AB}]$ equal 1 for all $r > 0$. This justifies the given interpretations of the adapted M index.

Jensen [2006, 2009] interpreted commercial activities that tend to aggregate as "friends", and such that tend to disperse as "enemies". In fact, Jensen [2009] argued that most existing stores are located at places that are "friendly" to them because badly situated stores would perish quite fast. In other words, the spatial distribution of the commercial activities seems to unravel interactions that favor or disfavor *successful* local coexistence of certain activities. From the spatial distribution of language features only, however, it cannot be directly determined which features *successfully* coexist and which do not, as at least half of the languages of the world are endangered and are therefore unlikely to be located in places that are friendly to them. Nevertheless, it seems

[1] Jensen [2009] explained that when measuring the local concentration of activity **A** around a store **s** of category *A* $(n_A(S)/n(S))$, it has to be compared to a reference concentration that does not take into account this particular store **s**. This reference concentration is obtained by subtracting 1 (for the store **s**) from the numerator and the denominator of the overall concentration of activity **A**. Thus, for every **A** store the reference concentration is $(N_A-1)/(N-1)$. In consequence, when averaging the ratio between the local concentration and the reference concentration over all **A** stores in the city, the latter concentration appears as a constant and can be factored out. When measuring the local concentration of a feature **A** in the neighborhood of a community **c** manifesting this feature *A* $(n_A(C)/n(C))$, in analogy, it has to be compared to a reference concentration that does not take into account this particular community **c**. This reference concentration is obtained by excluding the particular community **c** from consideration. In other words, this community's number of speakers $(n_{self}(C))$ has to be subtracted from the numerator and the denominator of the overall concentration of feature **A**. Therefore, for every feature **A** the reference concentration is $(N_A-n_{self}(C))/(N-n_{self}(C))$. In consequence, when averaging the ratio between the local concentration and the reference concentration over all communities **c** manifesting feature **A** in the entire sample area, the latter concentration does not appear as a constant like in Jensen's case (as $n_{self}(C)$ is variable), i.e. it cannot be factored out.

that in Eurasia, one can quantify interactions favoring or disfavoring *successful* coexistence between features by considering only communities that are probably not endangered when computing the M index, as they represent most of the speakers of the region. To this modification of the M index is referred to with *C index* in the following. More precisely the C index was defined as follows:

$$C_{AB} \equiv \begin{cases} M_{AA} & \text{for } A = B \\ M_{AB} & \text{for } A \neq B \end{cases} \quad (6)$$

where the considered linguistic communities are only the ones that are probably not endangered. The next section shows that quantifying this *coexistence ability* between features could be as much of interest as measuring their effective spatial aggregation or dispersion.

5 Predicting the vitality of minority languages

A location quality index for a feature, similar to the one Jensen [2009] had conceived for commercial activities, was defined:

$$Q_A(area) \equiv \frac{1}{\sum_{B \in F} n_B(area)} \sum_{B \in F} a_{AB} n_B(area) \quad (7)$$

where F is the set of all considered language features, $n_B(area)$ represents the sum of the number of speakers of the communities manifesting the feature B in the neighborhood of the given $area$ and a_{AB} is defined as follows:

$$a_{AB} \equiv \begin{cases} C_{AB} - 1 & \text{for } C_{AB} \geq 1 \\ -\left(\frac{1}{C_{AB}} - 1 \right) & \text{for } C_{AB} < 1 \end{cases} \quad (8)$$

The adapted Q_A index simply represents the average ability of a feature A, manifested by a community located on the given $area$, to coexist with the features of its neighboring communities (The transformation of C_{AB} allows to have disabilities to coexist – C_{AB} values inferior to 1 – on a same scale as positive abilities – C_{AB} superior to

1.). The location quality for a feature manifested by a community can be computed by removing the community from the sample and calculating Q_A for the *area* associated with this community.

The index is independent from the number of speakers in the neighborhood, i.e. independent from the number of network-entities around the *area*. This is the main difference to Jensen's index, which is amplified by the density of commercial activity in the neighborhood of *(x,y)*, i.e. amplified by number of network-entities around *(x,y)* (this can be verified mathematically by multiplying $n_B(x,y)$ of every point *(x,y)* by a same constant $k \in]0, \infty] \mid k \neq 1$). The reason for this difference is that for a linguistic minority community, a high population density in its neighborhood is susceptible to have a *normally rather negative* influence on its vitality: linguistic minority communities isolated on islands are in general less endangered than the ones situated on the continent [Sutherland, 2003]. As the influence of population density on vitality is not clear, the index was defined completely independent of this factor.

Based on this index the question whether and to what extent the linguistic structure of the communities in the neighborhood of a minority community allows predicting its level of endangerment can be illuminated: for each community its vitality can be compared to the localization qualities of the features it manifests (where a vitality can be assigned, of course). Due to the fact that the 31 considered communities share only few features and therefore also few predictors, for each community, the localization qualities of its features were aggregated (by averaging the z-scores) to form one single location quality index. These 31 community location quality indices were then put in relation to Unesco's corresponding vitality grades.

6 Results

The proposed M and C indices lead to results that are in agreement with visual verifications done. To give an example, figure 2 (next page) shows the spatial distribution of the communities manifesting the features having the ids 3686[1] and 760[2] in JM in the sample region (these features are not manifested in the omitted eastern part of the sample region). On this map, a strong spatial repulsion can be observed between these features. The computed M inter value is about 0.001, which means that in the neighborhood of the speakers manifesting the feature 3686, the average concentration of the speakers using the feature 760 is about a thousandth of the

[1] In JM described as "2.5.3.SIMPLE SENTENCE -> marginal constructions -> Affective"

[2] In JM described as "2.1.4.SYLLABLE -> the element following the vowel -> not more than one consonant"

overall concentration in the entire sample area. The visually observed repulsion seems coherently expressed by the computed M inter value.

Spearman's rang correlation between the (ordinal) vitality grades and the computed location qualities of the 31 considered minority communities has the value of 0.62 (p-value: 0.00009). In consequence, it seems that for the studied sample, the vitality of a minority language is indeed related to the linguistic structure of the neighboring languages. Besides, for all identified endangered communities, with the exception of five, the computed location quality Q_A is below zero, i.e. inferior to the one of an average community. On the other hand, three quarters of the communities that were judged to be probably not endangered (48), obtained a Q_A score above zero, i.e. superior to the one of an average community. It has to be noted though, that there is a certain circularity in the computation of the location quality indices for these latter communities as they had constituted the reference for the computations of the C_{AB} index on which Q_A is based. Figure 3 (next page) shows the score of the location quality computed for the above mentioned linguistic communities in function of their vitality. The 31 considered minority communities are represented as blue circles, the 48 communities that are probably not endangered as red squares.

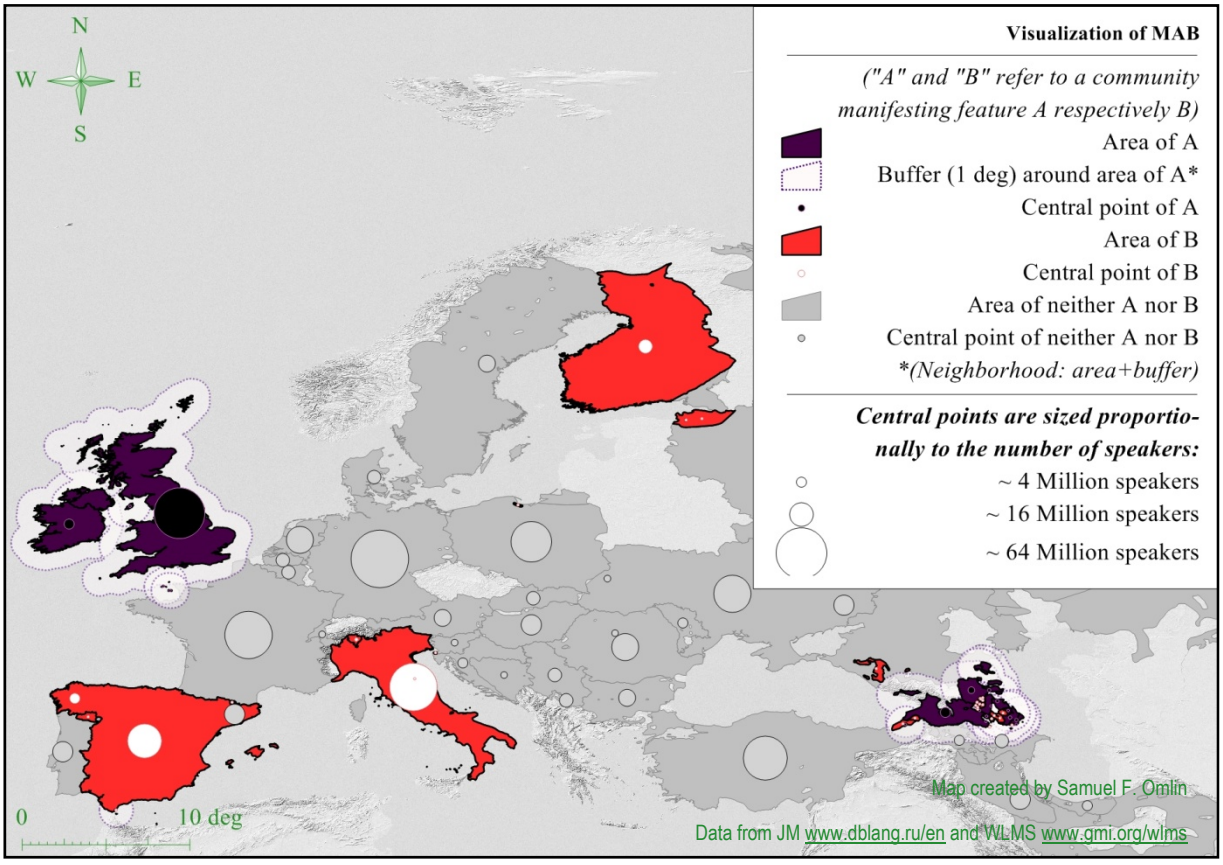

Figure 2: Visualization of M_{AB} computation for the features 3686 and 760.

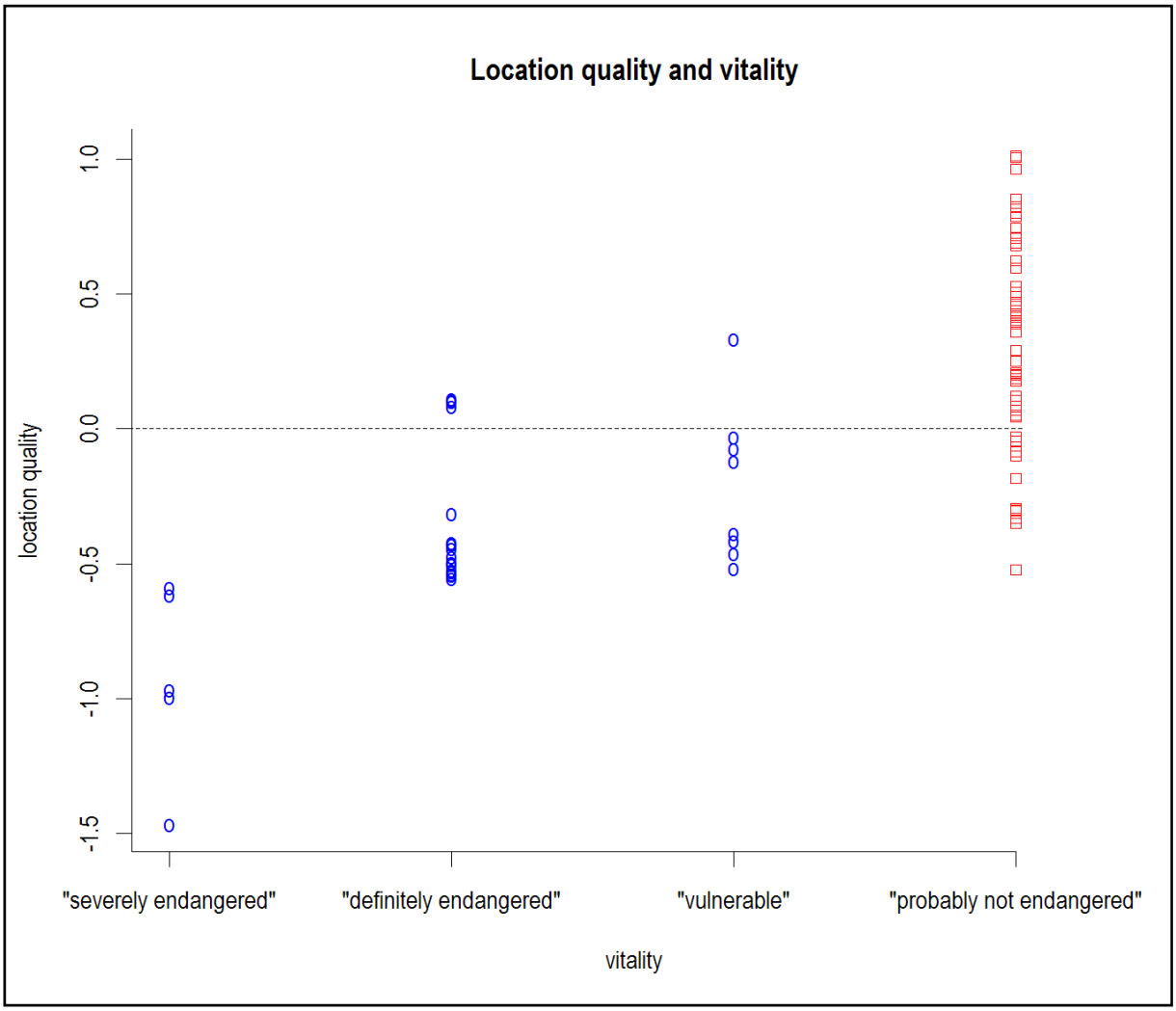

Figure 3: Location quality and vitality of linguistic communities.

Conclusions and future work

Studying the literature on the vast subject of endangered languages has allowed pointing out the importance of researching the influence of the geographic situation of a language on its survival and in particular the role and the importance of linguistic structures in this context. The main result obtained in this study confirms the relevance of the research questions raised: the degree of endangerment of the considered minority languages seems effectively related to the linguistic structure of their neighboring languages.

An approach has been proposed which allows estimating the importance of this relationship. This approach contains a method to measure the geographical distribution of linguistic features. Beyond the particular context of

predicting language vitality, this method constitutes a promising tool to unravel mechanisms that control the spatial distribution of language features.

The proposed approach was implemented with most current linguistic and geographical data: *Jaziky mira* [Academia and Indrik, 1993-2010], *World Language Mapping System* [GMI and SIL , 2010] and the *Atlas of the World's Languages in Danger* [Moseley, 2009].

Besides, it has been outlined how these three very recent databases can be joined in order to conduct quantitative linguistic studies when geographic parameters are involved.

Future work could contain the aspects listed in the following. Despite the fact that the presented methods seem to be quite robust to the buffer size, it would be valuable to study it. Further evaluation of the quality and reliability of the indices M and C would be beneficial. Professor Pablo Jensen's results from ongoing research could be of high value for this purpose[1]. As well, it should be researched whether the temporal stability of linguistic features plays a role for the measurement of their interactions. Following that, it may be promising to further explore alternatives to the complete aggregation of the location qualities of the features manifested by a linguistic community when predicting its vitality. Finally, to predict language vitality more precisely, further investigation of the role and relevance of language similarity in the context of language competition seems useful. A definition of language similarity based on typological features *which is adapted to the particular context of language competition* could constitute a helpful tool for this purpose.

Bibliography

[Abrams and Strogatz, 2003] D. M. Abrams and S. H. Strogatz. Linguistics: Modeling the dynamics of language death. Nature 424(6951), 900–900. 2003. [Online]. Available: http://dx.doi.org/10.1038/424900a

[Academia and Indrik, 1993-2010] Jazyki mira (Languages of the World). Moscow: Academia and Indrik, 1993-2010. [Online]. Available: http://ww.dblang.ru/en

[Castelló, Eguíluz, and Miguel, 2006] X. Castelló, V. M. Eguíluz, and M. San Miguel. Ordering dynamics with two non-excluding options: bilingualism in language competition. New J. Phys. 8(12), 308. 2006. [Online]. Available: http://dx.doi.org/10.1088/1367-2630/8/12/308

[1] Professor Jensen (French National Center for Scientific Research CNRS, France) has published some of his most recent results on his homepage: http://perso.ens-lyon.fr/pablo.jensen/.

[Castelló et al., 2007] X. Castelló, L. Loureiro-Porto, V. M. Eguíluz, and M. San Miguel. The Fate of Bilingualism in a Model of Language Competition. In: Advancing Social Simulation: The First World Congress, 83-94. Eds. S. Takahashi, D. Sallach, and J. Rouchier. Japan: Springer, 2007. [Online]. Available: http://dx.doi.org/10.1007/978-4-431-73167-2_9

[GMI and SIL, 2010] World Language Mapping System. Global Mapping International and SIL International, 2010. [Online]. Available: http://www.gmi.org/wlms

[Jensen, 2006] P. Jensen. Network-based predictions of retail store commercial categories and optimal locations. Phys. Rev. E 74(3), 035101(R). 2006. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.74.035101

[Jensen, 2009] P. Jensen. Analyzing the Localization of Retail Stores with Complex Systems Tools. In: Advances in Intelligent Data Analysis VIII: 8th International Symposium on Intelligent Data Analysis, Lecture Notes in Computer Science, 5772/2009, 10–20. Eds. N. M. Adams, C. Robardet, A. Siebes, and J-F. Boulicaut. Berlin Heidelberg: Springer-Verlag, 2009. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-03915-7_2

[Lewis, 2009] P. M. Lewis (Ed.). Ethnologue: Languages of the World, Sixteenth edition. Dallas, Texas: SIL International, 2009. [Online]. Available: http://www.ethnologue.com

[Mackey, 2001] W.F. Mackey. The ecology of language shift. In: The Ecolinguistics Reader: Language, Ecology and Environment, 67-74. Eds. A. Fill and P. Mühlhäusler. London: Continuum International Publishing Group, 2001.

[Mira and Paredes, 2005] J. Mira and A. Paredes. Interlinguistic similarity and language death dynamics. Europhys. Lett. 69(6), 1031–1034, 2005. [Online]. Available: http://dx.doi.org/10.1209/epl/i2004-10438-4

[Monteagudo and Santamarina, 1993] H. Monteagudo and A. Santamarina. Galician and Castilian in contact: historical, social, and linguistic aspects. In: Trends in romance linguistics and philology 5, 117–174. Eds. J. Green and R. Posner. Paris: Walter de Gruyter, 1993.

[Moseley, 2009] C. Moseley (Ed.). Atlas of the World's Languages in Danger. Unesco, 2009. [Online]. Available: http://www.unesco.org/culture/en/endangeredlanguages/atlas

[Palmer, 1997] S. Palmer. Language of work: the critical link between economic change and language shift. In: Teaching indigenous languages, 263-286. Ed. J. Reyhner. Flagstaff, Arizona: Northern Arizona University, 1997.

[Patriarca and Heinsalu, 2009] M. Patriarca and E. Heinsalu. Influence of geography on language competition. Physica A: Statistical Mechanics and its Applications 388(2-3), 174–186. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.physa.2008.09.034

[Patriarca and Leppänen, 2004] M. Patriarca and T. Leppänen. Modeling language competition. Physica A: Statistical Mechanics and its Applications 338(1-2), 296–299. 2004. [Online]. Available: http://dx.doi.org/10.1016/j.physa.2004.02.056

[Posner and Rogers, 1993] R. Posner and K. H. Rogers. Bilingualism and language conflict in Rhaeto-Romance. In: Trends in romance linguistics and philology 5, 117–174. Eds. J. Green and R. Posner. Paris: Walter de Gruyter, 1993.

[Schulze and Stauffer, 2006] C. Schulze and D. Stauffer. Monte Carlo simulation of survival for minority languages. In: Advances in Complex Systems (ACS) 9(3), 183–191. Ed. F. Schweitzer. Zurich: Frank Schweitzer, 2006. [Online]. Available: http://dx.doi.org/10.1142/S0219525906000719

[Sutherland, 2003] W. J. Sutherland. Parallel extinction risk and global distribution of languages and species. Nature 423(6937), 276–279. 2003. [Online]. Available: http://dx.doi.org/10.1038/nature01607

[Teşileanu and Meyer-Ortmanns, 2006] T. Teşileanu and H. Meyer-Ortmanns. Competition of Languages and Their Hamming Distance. International Journal of Modern Physics C 17(2), 259–278. 2006. [Online]. Available: http://dx.doi.org/10.1142/S0129183106008765

[Unesco, 2010] www.unesco.org/culture/en/endangeredlanguages. Safeguarding endangered languages. Unesco, 15.03.2010.

[Wang and Minett, 2005] W. S-Y. Wang and J. M. Minett. The invasion of language: emergence, change and death. Trends in Ecology & Evolution 20(5), 263–269. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.tree.2005.03.001

[Wardhaugh, 1987] R. Wardhaugh. Languages in competition: dominance, diversity, and decline. Oxford: Blackwell, 1987.

Author's Information

Samuel F. Omlin – *Section of Information Technologies and Mathematical Methods, University of Lausanne, CH-1015, Lausanne, Switzerland; e-mail: [firstname].[familyname]@unil.ch*

Major Fields of Scientific Research: Applications of Spatial Statistics, Spatial Analysis and Geographical Information Systems in Linguistics, Language Dynamics, Textual Statistics, Natural Language Processing, Supercomputing.

THE EXPERIENCE OF DEVELOPING SOFTWARE FOR TYPOLOGICAL DATABASES (ON THE EXAMPLE OF DB "LANGUAGES OF THE WORLD") [1]

Vladimir Polyakov

Abstract: In the present article we will discuss the experience of creating software for the typological database "Languages of the World". The DB "Languages of the World" is one of the biggest typological computer resources. We have done a review of the software connected with the DB "Languages of the World". The following questions are discussed: compatibility of the versions, choice of the best structure of the data, development of the content in newer versions of the DB, creation of bilingual version, correct citing. The main lessons learnt from the project by the workgroup, are:

Long development and creation of different versions of the product during its life cycle (over 20 years), providing its livability against the background of changing of operational systems and paradigms of programming makes us seriously think about a technology of providing for compatibility between different versions of the product, documenting of the code, preserving the key participants of the workgroup.

The structure of the DB is a secondary moment in the relation to the content. In the end, choice of a certain structure of data presentation in a certain realization of the product is a question of comfortable programming. Besides, choice of the structure of the data is in many situation defined by the environment of data storage, dates and budget of the product.

Planning a long life cycle of a linguistic resource for scientific purposes must foresee tools of fixation and archiving the inevitable changes of the content. Lack of such tools or links to the contents without invariant binding lowers the quality and the value of the received scientific results.

The creation of the bilingual version of the product demanded thorough elaboration of the terminological part of the DB, as well as linkage of the languages to the international system of coding. Along with it, the specificity of Russian scientific linguistic school and a more detailed description of the languages of Eurasia in the DB "Languages of the World" did not allow us to withdraw these contradictions completely.

The main scientific results received for the past 5 years with the use of the DB, are enumerated. The perspectives of its future development and use are studied.

Keywords*: language typology, linguistic database*

ACM Classification Keywords*: J.5 Arts and Humanities - Linguistics*

---

[1] The research was supported by Russian Scientific Foundation of Humanities (grant № 10-04-12125в)

1. Introduction

The success of informational technologies, abundance of linguistic information on different aspects of linguistics, presence of hard-to-solve scientific problems in this field promoted first the appearance of, first as shy attempts, and then wider spread of computer linguistic resources. Today using computer technologies in the sphere of linguistics is not rare, and it is even becoming a necessary element of both the process of studying and the process of learning. Computer technologies give the researcher and the teacher an incontestable competitive advantage.

We can consider the creation of text corpora for national languages a start point of using computers in linguistics [Francis & Kucera, 1967]. It promoted using quantitative methods (statistical, distributive) and later studies with the use of the technologies of the intellectual data analysis. The necessity of solving the problems of computer linguistics conditioned the growth of interest for the sphere of corpora researches, creation of representative corpora in different aspects of the theory and the practice of text and speech. The appearance of the net Internet makes these resources even more available, and allow to conduct researches in the remote regime [Bultreebank, Penn_treebank, PDT, EXMARaLDA]. Later researches on creating computer linguistic resources were conducted also in the sphere of linguistic semantics. The most striking examples of such resources are the projects WordNet [Miller, George A., 1995], [WordNet Online], Roget's Thesaurus [Roget, 1852], [Roget_Tes], FrameNet [Fillmore, Collin, 2010], [FrameNet], VerbNet [Kipper-Schuler, 2005], [VerbNet].

By the present time computer methods have been introduced into almost all spheres of linguistics. In the field of comparative linguistics lexical-statistical DB appeared historically first [Starostin]. Later there also appeared typological DB: Ethnologue - Languages of the World [Ethnologue], The World Atlas of Language Structures - WALS [Haspelmath et al., 2005], [WALS], UNESCO Atlas of the World's Languages in danger [Moseley, Christopher (ed.), 2010], Database "Languages of the World" of Institute of Linguistics of Russian Academy of Sciences [Polyakov, Solovyev, 2006], [DB_JM]. As opposed to researches in the sphere of corpus linguistics, which were first initiated by the problems of creating frequency dictionaries, and then by actual applied tasks (such as POS-tagging, morphological analysis, syntactic analysis, reference, etc.), in the linguistic typology the creation of DBs was initiated by the needs of fundamental scientific researches and partially by the educational sphere. A lot of discussion problems of typology, comparative linguistics, historical and areal linguistics, that had not been solved in the precomputer epoch, received a chance for solution in the conditions of large-scale use of databases and computer quantitative methods [Wichmann, Saunders, 2007], [Bayrasheva, Solovyev, 2008] [Solovyev, 2010].

In the present article we will discuss the experience of creating software for the typological DB "Languages of the World". The DB "Languages of the World" is one of the biggest typological computer resources. There are resources that exceed the DB in the number of covered languages, but so far there are no resources that have a comparable level of detail of description of such big language communities. The article [Polyakov, Solovyev, Wichmann, Belyaev, 2009] gives a detailed comparison of the DB "Languages of the World" and WALS. The project DB was started in the eighties of the XX-th century in Institute of Linguistics RAS and at the present time is positioned as an infrastructural scientific project. The DB has several versions, including Windows and Web-version. Around the kernel of the DB there were created a lot of programs for research purposes. Interesting

scientific results were received and new quantitative methods of scientific researched were worked out. The focus of the present article is on the problems that arise during the creation of software for such a long-term scientific computer project.

2. History of the project and a brief characteristic of the DB

In the 80-s in Institute of Linguistics of Russian Academy of Sciences (IL RAS) a decision about the launch of work on creation of the database Jazyki mira ("Languages of the World") was made. The encyclopedic issue of the same name [Jazyki mira, 1993…2006] is used as the source of information. The work was initiated by corresponding member of the Academy of Science Yartseva V.N. and was conducted at the department of applied linguistics under the surveillance of Novikov A.I. The following people took part in the development of the conception and the structure of the DB: Zotova A.K., Ryabtseva N.K., Rogova N., Romanova O.I. – analysis of abstracts, Vinogradov V.A., Zhurinskaya M.A., Testelets Ya.I., Yaroslavtseva E.I. – authors of the model, Skokan U.P., Novikov A.I., Nesterova N.N. – computer formalization of the model.

The first version of the DB was realized by programmer Skokan U.P. in DBMS Clipper (MS DOS). The registration certificate of Unitary Entreprise Scientific Technical Centre "Informregistr" № 7706 from November, 26, 2001 was received for the DB. A number of publications were made. In 2005 Yaroslavtseva E.I. defended a doctoral thesis on the topic "The computer database "Languages of the World" and its possible applications" [Yaroslavtseva, 2005].

In 2002 a Windows-version of the DB "Languages of the World" was created (project director – Polyakov V.N., programmer – Logunov V.). In 2005 the first variant of the Web-version was presented, and in March, 2006, the base was published on the Internet at www.dblang.ru (project director – Polyakov V.N., programmers – Goncharov E., Shcherbinin T., Khanukaev R.). A curriculum of the optional course "DB "Languages of the World" and new possibilities of typological and comparative researches" was worked out (authors: Polyakov V.N., Solovyev V.D.), it was read at the department of Theoretical and Applied Linguistics (philological department of Moscow State University) and at the department "Linguistics" of South Ural State University in 2006. In 2008 work on Reference and educational version of the DB with a more developed interface and new possibilities of search and navigation began (project director – Polyakov V.N., programmers – Belyaev O., Anisimov I.). This project is planned to summarize the experience gained for the past five years of researches and operation of the DB [Belyaev, 2008].

The work on creation and further development of the DB was repeatedly supported by grants of Russian Humanitarian Scientific Fond and Russian Fond of Fundamental Researches, partially is was financed by Moscow State Linguistic University, in 2006 and 2010 it became part of financing of Baudouin de Courtenay Russian Scientific-Educational Centre of Linguistics of Kazan State University (manager of scientific educational centre – Solovyev V.D.)

The Data Base "Languages of the World" (Jaziki Mira) has the following quantitative characteristics.

- contains more than 3800 features

- the number of languages is 313 Eurasian languages

- contains the description of the following spheres of language: phonetics, morphology, syntax.

- representation of data: binary

In Data Base "Languages of the World" the following language families and unities are represented: Austroasian, Austronesian, Altaic, Afroasian, Indoeuropean, Caucasian, Paleoasian, Sinotibetic, Uralic, Hurrito-Urartean. DB contains the description of languages-isolates: Ainu, Nivch, Burushaski, Sumeran, Elamite. The unique peculiarity of Data Base "Languages of the World" is a large collection of extinct languages description, that includes 55 essays. There is no analogues of such detailed and systematic description of extinct languages.

The main principles forming the model of language description are binarity, hierarchicity and paradigmaticity. The overall number of binary states in the DB is more than 1.2 million. In order to give an example of the complicity of processing such volume of data we will note that calculating a matrix of measures of similarity between all the languages on a modern personal computer takes more than 10 hours.

3. Review of software connected with the DB "Languages of the World"

Sources of Data for DB JM are:

- Encyclopedic issue "Jaziki Mira" (Languages of the World) – 15 volumes, printed by Institute of Linguistics of Russian Academy of Science from 1993 to 2009.
- Large Encyclopedic Dictionary. Linguistics (Edited by Yarceva V.N.) – includes interpretation of all terms of model of DB.

Main work on language description in DB format was fulfilled by Yelena Yaroslavceva, DSc.

The Kernel version includes full content of database and main functions for adding, editing and searching for data (see table 1).

About the conversion

Text abstracts were chosen as a way of converting the data during a transfer from the DOS-version to the Windows-version of the DB. During the direct data transfer on the level of DBF-files there appeared difficulties connected with the restrictions on the structure of DBF files and the coding of the Cyrillic alphabet. In the end transfer on the level of text abstracts turned out to be the best from the point of view of data safety. It required the support of two formats of text files in the Windows-version, because in the Russian Windows-version the coding Windows-1251 is used, and it is different from the coding of the Cyrillic alphabet CPP-866, which was adopted in the DOS-version.

The second problem that arose during the data transfer was connected with the necessity of replenishment the model of the abstract in the Windows-version. An abstract model is a list of features. In order to exclude possible

mistakes each abstract is checked for new features during the launch. If there are new features a specialist in the DB must introduce the necessary changes to the structure of the feature space or correct the mistake in the abstract. Then the abstract is loaded again. And it happens until there are no notices on mistakes. This technology ensures the high level of the control of reliability of the source data.

Table 1.

	Program language or environment / Database Engine and data format	Programmers, year of issue	English interface and content	Main functions	Compatibility
DOS Version	Clipper / Dbase compatible, DBF	Skokan †, 1997 (*)	Yes, but not synchronized with RUS-version on content	Correction of model, add new languages, browse, export, import, save, search , comparison	With Win version via files of essay export/import
Win Version	Pascal Delphi / Borland Database Engine, DBF	Logunov, Polyakov, 2002 (*)	Yes, but not synchronized with RUS-version on content	Correction of model, add new languages, browse, navigation, export, import, save, simple and complex search, comparison, alphabetic and thematic indices	With DOS version via files of essay export/import, with Web version via direct conversion of database files
Web version	C# and .NET / MS SQL Server	Goncharov (1st var.), 2005 Khanukaev (2nd var.), 2006 (**) There is also a Linux-version (at KSU).	The content Is fulfilled (Yaroslavtceva, Makarova). Interface is fulfilled (Khanukaev).	Browse, tree navigation, comparison	Loads data from Win version via direct conversion of database files

* Task formalization was done by Novikov †

** Task formalization was done by Polyakov.

Finally, certain difficulties arose also in connection with the creation of English version of the DB. It was historically established that works on the creation of the Russian and the English versions in the IL RAS were conducted separately. In the end the Russian version was far ahead of the English one in the number of languages. Now these lags are eliminated, but a decision to merge the two versions in one program product was made. It will eliminate the possibility of accidental mistakes and will lead to the unification of the space of languages and features.

It is necessary to note that works on coding the features in the DB require an exceptional mental outlook and patience. A specialist, who makes the coding, must be very good at the very difficult subject area in order to be able to establish binary equivalents of the features in the DB on the base of text descriptions. Moreover, a text of an encyclopedic article can contain hidden information that presupposes further explication or reference to a relative language. Yaroslavtseva E.I. did this work brilliantly. The examination of the DB conducted in the Kazan Center of Linguistics and IL RAS showed a relatively low level of mistakes. In future it is intended to create in the framework of the work group special questionnaires that allow to fill the abstracts with language descriptions in an easier way. An example of such questionnaire for section 2.1.1. PHONEMIC STRUCTURE was made by E. Loginova [Loginova, 2008].

About the structure

Such question as the data structure deserves a separate discussion. Initially the author of the first (DOS) version of the DB and the structure of the main data table Skokan made a decision to create a flat rectangular table, where rows in this table are features, and columns are languages (see pic. 1). In the field where the column (language) crossed the row (feature) there was the value True, if the given language had this grammar category, and False, if it did not. It is evitable that such approach does not conform to the canons of relational DBMS that are established in the modern ITs. But at that moment – mid 80-s – such data format, taking into consideration the volume of data, available computer powers, was optimal from the point of view of efficiency, speed of search and data retrieval.

	Language 1	Language 2	…	Language m
Feature 1	True	True	…	False
Feature 2	True	False	…	False
…	…	…	…	…
Feature n	False	True	…	True

Pic. 1. Structure of the main table of data

When the data were transferred from the DOS-version to the Windows-version, it was decided to keep the stated data structure. But in the latest Web-version this structure was transformed into a relational format for supporting SQL-queries. The relational format of the DB was also adopted in the work [Omlin, 2010]. Special programs-converters were created in order to transfer the data from the table to the relational structure.

About the content

Initially the content of the DB was limited to the contents of the encyclopedia articles. At the same time some sections of articles (mainly sociolinguistical and ethnographical) were loaded to the MEMO-field and were unavailable for search. All other sections were coded in the binary system of values. During the shift to the educational-inquiry version it was decided to eliminate this shortcoming. Besides, a number of data arrays that present interest for query forming, were added to the DB.

The following data were added to the educational-inquiry version of the DB:

-text of a language description from the encyclopedia in PDF-format (only in Russian);

-glossary;

-genetic index;

-references to literature sources;

-examples demonstrating the meaning of grammar categories;

-status of the language (literature, education, writing);

-changes due to contacts (grammatical and lexical);

-number of speakers;

-frequency of feature spread in macro-families, families, branches, groups, subgroups;

-geographic coordinates [Loginova, 2009];

-the language code according to the standard ISO 639-3 (see www.ethnologue.com);

-for most features there is a link to the corresponding article in the online encyclopedia Wikipedia.

Moreover, every quantitative program product creates its own digital content, which is difficult for transferring to the main DB. These data are usually published by the researchers separately.

About the English version

During the creation of the English version there were found terminological equivalents for languages and grammar categories. Sometimes it was hard to do. For example, in the DB there are dialects of the languages whose names are neither at the web-site www.ethnologue.com nor in the encyclopedia Wikipedia. In this case we had to make a calque from the Russian language. There were similar problems when grammar features were translated.

About the versions of the DB

One of the problems that we cannot manage so far is fixing and operating the archive of changes in the content of the DB. Changes in the DB can be caused by an expertise of existing language descriptions or adding new language descriptions. Due to this some quantitative calculations made in previous versions of the content, can give a wrong link in a newer version during referring to the numbers of features. That is why in publications it is necessary to give the date of calculations and the version of the DB. Besides, it is desirable to specify the full path of a certain grammar category in the tree of features when a reference to this feature is given.

For example:

Nominative/accusative<=subjective-objective <=argument case meanings <=2.3.4.CASE MEANINGS (1359)

Here the number in brackets is the identifier of the feature in the DB, the tree root is in the right part of the line.

For the past five years of researches connected with the DB, the work group have worked out several program products.

In table 2 different kinds of software products related to DB JM are represented.

The last row of the table (Outer tools applicable to JM data) presents the products of detached developers that are used during the work with the contents of the DB.

Quantitative and other research products connected with DB JM are described in table 3.

The apocryphal structure of the main table turned out to be very convenient during the transfer of the data to the format MS Excel, and the built-in language VBA allowed to create a number of successful quantitative products (Similarity, LangFam). Thus, today the DB is used in three data formats: DBF, XLS, SQL Server.

Also some referential tools are developed (see table 4).

It is possible that in future there will appear new specialized products containing a fragment of the DB and supplemented with some new information. There already exists such practice. For example, in the project [Omlin, 2010] the content of the DB "Languages of the World" was united with the content of the project UNESCO (Atlas of the World's Languages in Danger) [Moseley, Christopher (ed.), 2010]. There are joint works with New Bulgarian University on creating a specialized version of the DB that will be dedicated to the case system of languages.

Table 2.

Versions of Database	DOS Version	Windows Version	Web Version www.dblang.ru
Quantitative And Other Research Products	Includes comparison of two languages as function	Similarity – Software for similarity measure calculations LangFam – Software for language family portraits calculations, genetic markers revealing, deal with rare features filters, investigate typological shift etc. Special software for modeling of evolution Special software for clusterization task Special software for phylogeny with different metrics of feature space BiCoTree –software for easy tree building on DB. Some other research programs, developed for different aims during partial investigations in areal, historical and typological linguistics (Gusareva, Loginova, Fashutdinov, Omlin, Polyakov, Solovyev).	Includes comparison of two languages as function
Reference and Educational Products (under constr.)		Living Diagrams – reference software with possibility of integration source data and quantitative diagrams EduDBLANG – educational version of DB with full spectrum of reference possibilities	The Web-version of "Living diagrams" is prepared.
Outer tools applicable to JM data		R – statistical sotware tools Different phylogeny tools.	

Table 3.

Product	Program language or environment / Database Engine and data format	Programmers, year of issue	Main functions
Similarity	VBA, Excel	Polyakov, 2006	Similarity measure calculations and evaluation
LangFam	VBA, Excel	Polyakov, 2006	Software for language family portraits calculations, genetic markers revealing, deal with rare features filters, investigate typologycal shift etc.
Special software for modeling of evolution	Pascal Delphi	Yuzhikov, 2006 (*)	Modeling of process of appearance, borrowing, extinction of features. Uses different parameters of model, gives different quantitative values.
Special software for clusterization task	Pascal Delphi	Dvoenosova (1st var), 2006 Zheleznovsky (2nd var), 2008 (*)	Clusterization of languages and features by different techniques of classic cluster analysis
Special software for phylogeny wspaceith different metrics of feature	Visual C	Faskhutdinov , 2008 (*)	Use two heuristic ideas of L- and S-metrics for calculation of distance between languages.
BiCoTree –software for easy tree building on DB.	Pascal Delphi	Sarvarov, 2010 (*)	
Some other research programs, developed for different aims during partial investigations in areal, historical and typological linguistics	C, Pascal Delphi	Gusareva, Loginova, Fashutdinov, Omlin, Polyakov, Solovyev	Allow to solve different tasks: - To calculate a core of relevant features for different language families; - To calculate a motherland for different language families using grammar features; - To calculate stability index using different metrics; - Etc.

Table 4.

Product	Program language or environment / Database Engine and data format	Programmers	Main functions
Living Diagrams	C# and .NET MS SQL Server Excel	Khanukaev (*)	Reference software with possibility of integration source data and quantitative diagrams. Allows to draw quantitative pictures or tables and to do queries to source data immediately from picture. Has purpose to improve confidence of linguists to quantitative results.
EduDBLANG	C# and .NET MS SQL Server Excel	Belyaev (*)	Educational version of DB with full spectrum of reference possibilities. Includes genetic and geographic indices, annotation and examples for features, full texts of papers according to the best WALS traditions. New concept of user interface.

4. Lessons, prospects, scientific results

Let us formulate the main lessons that were learnt by the work group from this project:

Long development and creation of different versions of the product during its life cycle (over 20 years), providing its livability against the background of changing of operational systems and paradigms of programming makes us seriously think about a technology of providing for compatibility between different versions of the product, documenting of the code, preserving the key participants of the workgroup.

The structure of the DB is a secondary moment in the relation to the content. In the end, choice of a certain structure of data presentation in a certain realization of the product is a question of comfortable programming. Besides, choice of the structure of the data is in many situations defined by the environment of data storage, dates and budget of the product.

Planning a long life cycle of a linguistic resource for scientific purposes must foresee tools of fixation and archiving the inevitable changes of the contents. Lack of such tools or links to the contents without invariant binding lowers the quality and the value of the received scientific results.

The creation of the bilingual version of the product demanded thorough elaboration of the terminological part of the DB, as well as linkage of the languages to the international system of coding. Along with it, the specificity of Russian scientific linguistic school and a more detailed description of the languages of Eurasia in the DB "Languages of the World" did not allow us to withdraw these contradictions completely.

Let us enumerate the scientific results received for the five past years of using the DB in scientific researches.

- On the base of the data, a new quantitative model of language evolution was worked out and approved by V.D. Solovyev. The use of this model allowed to receive an invariant curve, which reflects the stages of evolution: diagram "Language-Feature" (LF-diagram) [Polyakov, Solovyev, 2006].

- Polyakov V.N. received data on a new diachronic phenomenon, which the author called Typological Shift [Polyakov, Solovyev, 2006], [Polyakov, Yaroslavtseva, 2008]

- Polyakov V.N. worked out and optimized algorithms for calculating measures of language similarity on the base of the similarity of their grammar structure. For the first time a good coincidence of the values of likeness and the data on genetic relationship of languages was shown [Polyakov, Solovyev, 2006] [Polyakov, 2008], and it allowed to bring forward a thesis that the grammar structure carries sufficient information on genetic relationship, which was earlier questioned.

- In the period from 2006 to 2010 a group under the guidance of Solovyev V.D. has made a valuable contribution to adaption of methods of phylogeny to calculation of genetic trees on the base of the grammar structure of languages. New metrics and algorithms of calculating genetic trees were introduced. [Solovyev, 2007], [Solovyev, Fashutdinov, 2008], [Solovyev, Fashutdinov, 2009], [Solovyev, 2009]

- Polyakov V.N. introduced and improved a number of methods for formulating and verifying genetic hypotheses and/or areal contacts, including: method of ranging languages according to the value of similarity measure, method of quantitative maps of language communities, method of quantitative filters, method of filters according to genetic markers [Polyakov, Solovyev, 2006].

- Group under the guidance of Polyakov V.N. revealed genetic markers for the Altai language family, revealed the core of features for the Ural languages (in print).

- Parameters of the age of language families of Eurasia are specified [Solovyev, Fashutdinov, 2009-2], [Solovyev, 2009-2], [Bayrasheva, Solovyev, 2008-2].

- In cooperation with colleagues from Max Planck Institute for Evolutionary Anthropology indices of stability of grammar categories are calculated [Wichmann, Holman, 2009], [Belyaev, 2009], [Solovyev, Fashutdinov, 2009-3];

- A number of new methods and results of quantitative calculations for comparative linguistics, typology and areal linguistics are worked out (in print).

In future the development of the DB will be conducted in the following directions:

- development of new quantitative products and joint DBs for solving the problems in the sphere of linguistic typology, comparative linguistics, historical and areal linguistics;

- generalization of separate quantitative products and technologies in the framework of a united technological environment;

- integration of numerical data of quantitative calculations in the united DB;

- creation of geographic information applications with the use of the contents of the DB.

5. Conclusions

The DB "Languages of the World" is a linguistic resource that has value for the sphere of researches and education. The history of elaboration and development of the DB, software connected with it, researches on its basis, allowed to gain important experience of conducting such large-scale interdisciplinary and interinstitutional projects. At the present time the DB is an infrastructural scientific resource that has a high scientific potential, involved in the international scientific society and providing a few levels of use in the sphere of science and education.

Bibliography

[Bayrasheva, Solovyev, 2008] Solovyev V., Bayrasheva V. Statistic analysis of linguistic databases: the new perspective in the typology and comparative studies. Text processing and cognitive technologies. V.17. 2008, p. 210 – 214.

[Bayrasheva, Solovyev, 2008-2] Bayrasheva V., Solovyev V. Modelling the Evolution of Language Features. Proc. of the intern. conf. "Cognitive and Functional Perspectives on Dynamic Tendencies in Languages". Tartu: University of Tartu. 2008. p. 198-199.

[Belyaev, 2008] Oleg Belyaev. A New Interface Model of the DB "Languages of the World". In Text Processing and Cognitive Technologies. Paper Collection. (Edited by V. Solovyev, M. Bergelson, V. Polyakov). The X-th International Conference "Cognitive Modeling in Linguistics". Proceedings. Volume 3. Kazan: KSU, 2008, p. 118-128.

[Belyaev, 2009] Belyaev, Oleg. 2009. Stability of language features: a comparison of the WALS and JM typological databases. Paper presented at Cognitive Modeling in Linguistics–2008, September 6–12, Bechichi, Montenegro. In print. Available online at: http://obelyaev.googlepages.com/BelyaevJMStab.pdf .

[Bultreebank] HPSG-based Syntactic Treebank of Bulgarian. URL: http://www.bultreebank.org/

[DB_JM] Database "Languages of the World". Institute of Linguistics. Russian Academy of Sciences. URL: http://www.dblang.ru/en/Default.aspx

[Ethnologue] Ethnologue, Languages of the World. An encyclopedic reference work cataloging all of the world's 6,909 known living languages. URL: http://www.ethnologue.com/

[EXMARaLDA] EXMARaLDA: SFB 538 Corpora. Spoken Language Corpora at the Research Center on Multilingualism. URL: http://www.exmaralda.org/corpora/en_sfbkorpora.html

[Fillmore, Collin, 2010] Fillmore, Charles J. & Baker, Collin F. 2010. A Frame Approach to Semantic Analysis, in Heine, B. & Narrog, H. (eds.) Oxford Handbook of Linguistic Analysis

[FrameNet] FrameNet Project.

URL: http://framenet.icsi.berkeley.edu/index.php?option=com_frontpage&Itemid=1

[Francis & Kucera, 1967] Francis S. and Kucera H., Computational analisys of present-day American English, Providence, RI: Brown University Press, 1967.

[Haspelmath et al., 2005] Haspelmath, Martin & Matthew S. Dryer & David Gil & Bernard Comrie (eds.) The World Atlas of Language Structures. − Oxford: Oxford University Press, 2005. − 695 p.

[Jazyki mira, 1993] Jazyki mira: Ural'skie jazyki (Languages of the World: Uralic languages). 1993. Moscow.

[Jazyki mira, 1997] Jazyki mira: Tûrkskie jazyki (Languages of the World: Turkic languages). 1997. Moscow: Indrik.

[Jazyki mira, 1996] Jazyki mira. Paleoaziatskie jazyki (Languages of the World. Palaeoasiatic languages). 1996. Moscow: Indrik.

[Jazyki mira, 1997] Jazyki mira: Mongol'skie jazyki. Tunguso-Man'čžurskie jazyki, Japonskij jazyk. Korejskij jazyk. (Languages of the World: Tunguso-Manchurian languages. Japanese language. Korean language). 1997. Moscow. Indrik.

[Jazyki mira, 1997] Jazyki mira: Iranskie jazyki. I. Jugo-zapadnye iranskie jazyki (Languages of the World: Iranian languages. I. Southwest Iranian languages). 1997. Moscow: Indrik.

[Jazyki mira, 1998] Jazyki mira: Dardskie i nuristanskie jazyki (Languages of the World: Dardic and Nuristani languages). 1998. Moscow: Indrik.

[Jazyki mira, 1999] Jazyki mira: Germanskie jazyki. Kel'tskie jazyki (Languages of the World: Germanic languages. Celtic languages). 1999. Moscow: Academia.

[Jazyki mira, 1999] Jazyki mira: Iranskie jazyki. II. Severo-zapadnye iranskie jazyki (Languages of the World: Iranian languages. II. Northwest Iranian languages). 1999. Moscow: Indrik.

[Jazyki mira, 1999] Jazyki mira: Iranskie jazyki. III. Vostočnoiranskie jazyki (Languages of the World: Iranian languages. III. East Iranian languages). 1999. Moscow: Indrik.

[Jazyki mira, 2001] Jazyki mira: Kavkazskie jazyki (Languages of the World: Caucasian languages). 2001. Moscow: Academia.

[Jazyki mira, 2001] Jazyki mira: Romanskie jazyki (Languages of the World: Romance languages). 2001. Moscow: Academia.

[Jazyki mira, 2004] Jazyki mira: Indoarijskie jazyki drevnego i srednego perioda (Languages of the World: Old and Middle IndoAryan languages). 2004. Moscow: Academia.

[Jazyki mira, 2005] Moldovan, A. M., S. S. Skorvid, A. A. Kibrik et al. (eds.). 2005. Jazyki mira: Slavânskie jazyki (Languages of the World: Slavic languages). Moscow: Academia.

[Kipper-Schuler, 2005] Karin Kipper-Schuler. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, University of Pennsylvania.

[Loginova, 2008] Liza Loginova. The Problems of Representation of Typological Data About Language in the Format of the Database (On the Material of the DB "Languages of the World") In Text Processing and Cognitive Technologies. Paper Collection. (Edited by V. Solovyev, M. Bergelson, V. Polyakov). The X-th International Conference "Cognitive Modeling in Linguistics"(CML-2008) . Proceedings. Volume 3. Kazan: KSU, 2008, p. 188-195.

[Loginova, 2009] Elizaveta Loginova. Technique of Definition of Geographical Coordinates at the Identification of the Area of Distribution of Language (On the Material of DB «Jaziki Mira»). In Text Processing and Cognitive Technologies. Paper Collection. The XI-th International Conference "Cognitive Modeling in Linguistics" (CML-2009). Proceedings. Kazan: KSU, 2009.

[Miller, George A., 1995] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[Moseley, Christopher (ed.), 2010] Moseley, Christopher (ed.). 2010. Atlas of the World's Languages in Danger, 3rd edn. Paris, UNESCO Publishing. Online version URL: http://www.unesco.org/culture/languages-atlas/

[Omlin, 2010] Samuel Omlin. Study of the Relation of the Linguistic Environment of a Minority Language to Its Vitality. In Text Processing and Cognitive Technologies. Paper Collection. The XII-th International Conference "Cognitive Modeling in Linguistics"(CML-2010). Proceedings. Kazan: KSU, 2010.

[PDT] The Prague Dependency Treebank 1.0. URL: http://ufal.mff.cuni.cz/pdt/

[Penn_treebank] Penn Treebank Project. URL: http://www.cis.upenn.edu/~treebank/

[Polyakov, 2008] Polyakov V. Approaches to improvement of similarity measure based on the structure of language description in the DB "Languages of the World". Text processing and cognitive technologies. v.17. 2008, p. 192 – 209.

[Polyakov, Solovyev, 2006] Polyakov, Vladimir N. and Valery D. Solovyev. 2006. Komp'juternye modeli I metody v tipologii i komparativistike (Computational Models and Methods in Typology and Comparative Linguistics). Kazan: Kazanskiy Gosudarstvennyy Universitet. 208 p.

[Polyakov, Solovyev, Wichmann, Belyaev, 2009] Polyakov V., Solovyev V., Wichmann S., Belyaev O. Using WALS and Jazyki mira. Linguistic Typology. V. 13. 2009. P. 135–165.

[Polyakov, Yaroslavtseva, 2008] Polyakov V.N., Yaroslavtseva E.I. Kvantitativnie zakonomernosti tipologicheskogo sdviga v yazikah Evrazii (The Quantitative Parameters of Typological Shift in Languages of Eurasia). Uchenie zapiski KGU, Vol.150, Book 2, 2008, p. 97-118.

[Roget, 1852] Roget, Peter Mark [1852] (1962), Dutch, Robert A., O.B.E., ed., The Original Roget's Thesaurus of English Words and Phrases (Americanized edition), New York, NY, USA: Longmans, Green & Co./Dell Publishing Co., Inc.

[Roget_Tes] ROGET's Hyperlinked Thesaurus. URL: http://www.roget.org/index.htm

[Solovyev, 2007] Solovyev V.D. Zadachi i metodi lingvisticheskoy filogenomiki (The tasks and the Methods of Linguistic Phylogeny). Conference "Znaniya, ontologii, teorii". Proceedings. Novosibirsk. Siberian Department of RAS, 2007.

[Solovyev, 2009-2] V.D. Solovyev. Viyavlenie sluchaev parallel'noy evolyucii s pomoshch'yu bazi dannih "Yaziki mira"(Revealing of Cases of Parallel Evolution by Means of a Database "Languages of the World") The 8th Conference on Languages of Far East, Southeast Asia and West Africa (LESEWA-8). Proceedings. Moscow. 2009.

[Solovyev, 2009] Solovyev V.D. Problemi i metodi lingvisticheskoy filogenii (Problems and Methods of Linguistic Phylogeny). Uchenie zapiski KGU. Vol. 151, Book 6, 2009. P. 8-21..

[Solovyev, 2010] Solovyev V.D. Tipologicheskie bazi dannih: perspektivi ispol'zovaniya (Typological databases: prospects of usage). Voprosi yazikoznaniya, 2010, №1. p. 94-110

[Solovyev, Fashutdinov, 2008] V.D. Solovyev, R.F. Fashutdinov. Vibor metriki dlya filogeneticheskih algoritmov (Choice of the Metrics for Phylogenetic Algorithms). Scientific Session of Moscow Engineering Physics Institute. Proceedings, Vol. 10. Moscow: MEPhI, 2008. p. 176.

[Solovyev, Fashutdinov, 2009] V.D. Solovyev, R.F. Fashutdinov. Preobrazovanie metrik, ispol'zuemih v metodah klasterizacii dlya postroeniya filogeneticheskih derev'ev yazikov (Transformation of the Metrics Used in Methods of Clusterization for Construction of Phylogenetic Trees of Languages). Uchenie zapiski KGU. Vol. 151, Book 3. 2009. P. 229–239

[Solovyev, Fashutdinov, 2009-2] V.D. Solovyev, R.F. Fashutdinov. Metodika kolichestvennoy ocenki skorosti evolyucii grammatiki (Technique of a Quantitative Estimation of Speed of Evolution of Grammar). Scientific Session of Moscow Engineering Physics Institute. Proceedings, Vol. 4. Moscow: MEPhI, 2009.

[Solovyev, Fashutdinov, 2009-3] V.D. Solovyev, R.F. Fashutdinov. Metodika ocenki stabil'nosti grammaticheskih svoystv (Technique of an estimation of stability of grammatical properties). Izvestiya RAN. Seriya literaturi i yazika. Vol.68. № 4. 2009.

[Starostin_DB] An Etymological Database Project. URL: http://starling.rinet.ru/main.html

[VerbNet] VerbNet Project. URL: http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[WALS] The World Atlas of Language Structures Online. URL: http://wals.info/

[Wichmann, Holman, 2009] Wichmann, Søren and Eric W. Holman. 2009. Assessing Temporal Stability for Linguistic Typological Features. München: LINCOM Europa. 82 p.

[Wichmann, Saunders, 2007] Wichmann, Søren and Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. Diachronica 24.2: 373-404.

[WordNet Online] Princeton University "About WordNet." WordNet. Princeton University. 2010. URL: http://wordnet.princeton.edu

[Yaroslavtseva, 2005] Yaroslavtseva E.I. Komp'yuternaya baza dannih "Jaziki Mira" i ee vozmozhnie primeneniya (Computer database « Languages of the World » and its possible applications). Dr.of S. Thes. – IL RAS, 2005.

Authors' Information

Vladimir Polyakov *– Senior Researcher at the Department of Applied Linguistics of Institute of Linguistics of Russian Academy of Sciences, Assoc. Professor of National Research Technological University (MISIS) and Moscow State Linguistic Universit, PhD.*

Address: Leninsky Avenu, 4. Moscow. Russia; e-mail: pvn-65@mail.ru

Chair of Organizing Committee of International Conference "Cognitive Modeling in Linguistics.

Major Fields of Scientific Research: Cognitive Science and Modeling, Computer Linguistics, Artificial Intelligence.

Projects: www.dblang.ru ; www.cml.msisa.ru ; www.finforecast.ru

MODEL RESEARCH OF INTERACTION PROCESSES OF TEXT SPACES

Konstantin I. Belousov, Tatyana N. Galinskaya

Abstract: The article discusses the problem of interaction of text spaces. When discussing the interaction of text spaces we assume that there exists a certain text model. The technique of semantic charting and the method of positional analysis allowed us to represent the successive-simultaneous semantic space of a text as its "semantic outline". Owing to the method of the prosodic analysis of a text, aimed at modeling its prosodic outline, there appears the possibility to analyze the cooperative interactions of these relatively independent text spaces. The system-approached research program presented in the work is aimed at the study of the text as a polyontological, self-organizing spatiotemporal linguistic object. The multiaspect text analysis is grounded on a) the positional analysis method, b) quantitative methods which in there turn comprise such methods as c) correlation methods, which determine the text aspects' level. By comparing and contrasting synchronically semantic connection intensity and mean sound intensity of the obtained data we received the results that allow us to be more specific in the discussion of the text structure as an evolving process. The search for explanatory tools of convergence, divergence, intersection, overlapping of various text structures is the key to understanding the complex material, ideal and social nature of text, its presentation as wholeness.

Keywords: *system activity approach, modeling, positional analysis, semantic charts, semantic graph of a text.*

Introduction

The concept "spaces" of the text is common to many works studying the text. However, it is used rather figuratively as an opposition of a certain (closed) reality to another reality (cf, e.g. the semantic space of a text, the emotive space of a text, etc). Due to the metaphorical character of the term, it remains unclear how "spaces of the text" are related to the text and to any other of its spaces. Thus the basic notion – the text – is ambiguously defined. The multiaspect text analysis is grounded on a) the positional analysis method, b) quantitative methods which in there turn comprise such methods as c) correlation methods, which determine the text aspects' level.

Studying the spaces of a text we usually start from the idea of a text as a functional system, hence, a text is usually regarded as being focused on "achieving certain goals, accomplishing a certain extralinguistic task" [Leont´ev, 2001]. Being a linguistic product by nature a text is created for pragmatic purposes and realized in the

field of pragmatics. Such understanding of a text is quite pertinent, but much narrower than the notion of the functional system as seen by P. K. Anochin [Anochin, 1999]. According to the theory of functional systems, developed by Anochin, the useful result (the pragmatic purpose in terms of linguistics) gains the status of a factor which causes the individual components to interact "in accordance with the type of their cooperation." [Anochin, 1999]. However, the study of the mechanisms controlling the interaction of the components (or subsystems) by which the favorable result for the whole system is gained, is skipped in the research dealing with the functionality of a text. Thus functionality becomes another linguistic metaphor. Hence it seems reasonable to introduce the prime statements regarding the subject of inquiry. We understand the text as an integral polyontological linguistic object which exists in forms of space-time, thus the main attributes of the text are:

1) Existence in space-time reality. The text entirely displays its existence abiding by the basic patterns of matter motion, and its content is also revealed according to these patterns. Unlike the meaning of a sign the text content extends in space-time since it is developed by the purposefully organized chronotheme sequence of techniques, actions and operations, united by various methods and approaches to the text.

2) We consider text as a phenomenon having various levels of existence (various ontologies) that lie between the two poles: material (to which physical attributes of a text, disclosing its acoustic-wave ontology) and conceptual.

3) The text organization on different ontological levels has common features as far as the text "accomplishes" extralinguistic tasks. This fact makes it necessary to introduce the concept of cooperation of text spaces. The analysis and the subsequent synthesis of text spaces appear feasible owing to the system activity approach, the method of positional text analysis as well as theoretical constructs (form, structure, etc.).

In compliance with goals and problems pursued, every studied *object* occupies a certain *object domain* (aspect) which can be realized independently. There can be several domains of this kind. By means of classification, abstraction, analysis techniques and synthetic procedures every object domain appears as a system of hierarchically arranged elements connected with each other by relations of various nature. Therefore, referring to the system/structure, we mean the system/structure of a certain aspect, not the whole object. First it is necessary to single out an aspect in the given object before carrying out its system analysis. And the focusing of the aspect is in the basis of the activity approach. Thus combining the system and activity approaches is inevitable when studying the objects of the objective reality. That is why we should not use the term "system object", but the term "system aspect", which implies the integrity of both the system and activity approaches.

Positional Analysis Method of the Text

The essence of the positional analysis is in the marking of the language units in a linear sequence. As long as a text is limited extensively (has the beginning and the end) the beginning is taken as "0", the end as "1", irrespective of the size of a text. Owing to this convention we can compare texts of diverse sizes as well as results of multiaspect descriptions of the same text. All language elements of a given text are positioned linearly in a row. The word is considered a counting unit. In order to locate an element, one should define the position of a word containing a required element (in case segments are smaller than a word) or coinciding with it (in case the word "level" is studied) [0; 1]. Then a simple arithmetic operation is performed, that is dividing the ordinal number of the pertinent word, by the total number of words in the text. The research based on the principles of the system and activity approach comprises the following steps:

1) singling out certain aspects within an object and presenting them as abstract systems;

2) the quantitative description of them including the spatiotemporal order, indicating the appearance and functioning of a system aspect element (description by means of the method of positional analysis). The description itself is the study of an attribute´s appearance intensity (or probability) in every point of the temporal development of the text;

3) converting *absolute* number values of the process rate in a certain spatiotemporal position into *relative* number values, i.e. the values within the [0; 1] interval (in order to compare description results for different aspects). This operation is performed for all text aspects. If we present the description results graphically, we obtain the "outline" of the text aspects (semantic, prosodic, etc); [we understand the term 'outline' as graphic representation of text aspects (dynamic intensity in text unfolding - see Figure 3)].

4) By means of correlation methods of statistics (the Pearson correlation method in particular) and methods of graphical representation the obtained constructs from different text aspects are systematized due to their general ontological basis. Graphical representation allows to compare similarly modeled text outlines.

Thus the research strategy comprises a number of operations such as idealization, modeling, object abstraction. Besides, one of the major operations is the statistic analysis. The correlation methods of this analysis is a tool to integrate multifarious research. In this respect the status of correlation as a research method becomes philosophically important, since correlation acts as integration means for various aspects of one object.

Form of a Text and Text Structures

For the above described procedure of text aspect integration, the concept "form of a text" is suggested. Every aspect is manifested as a system by means of abstracting, analysis and modeling and is "endowed" with its own structure. Undoubtedly, structures of individual aspects may be different. However, these structures should have some common features as these are aspect structures of one and the same object. *This commonalty is a form which is inherent in the object.* While structure is a derivative from the research activity – that is – its construct, the form is the phenomenon. When we study the form of a text in its certain aspects it transforms into its structure. Thus we deal with *form projection on the aspect domain*. That's why form can be reconstructed by comparing its constituent structures. We point out general and occasional in its structures, which is not possible in case of monoaspect object description (See Figure 1).

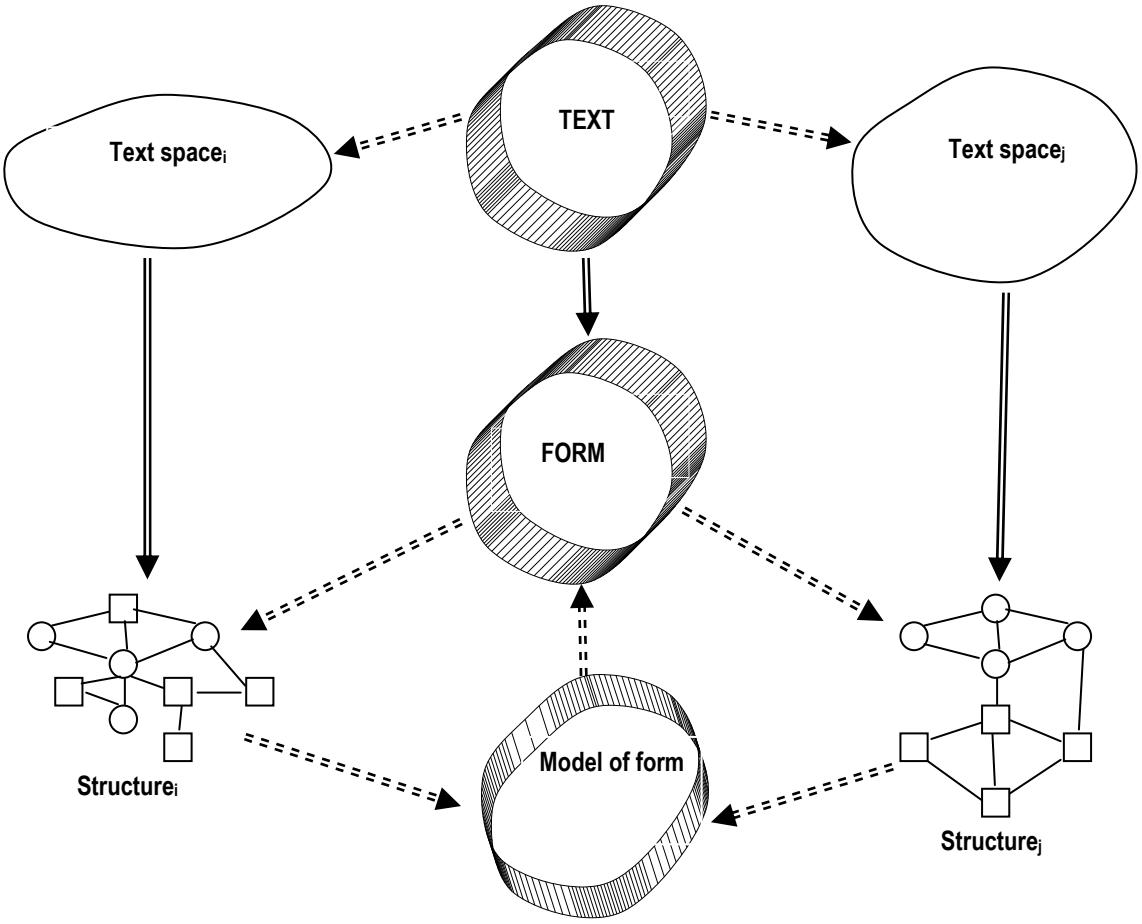

Figure 1. Form of the text and its structures

Semantic Charting of a Text

The study of the semantic text organization was carried out with the help of semantic charting of text – the method which has been elaborated by the author of the article. The semantic text charting is carried out according to the results of text assignments, performed by informants. The informants (21 philological faculty students) are given the text and assigned 1) to read the text and to define its theme; 2) to identify microthemes of the text; 3) to distribute the words of the text into semantic groups according to the identified microthemes, 4) each group should be entitled (the informants are expected to explain the grouping principle; the number of groups and words referred to them being unlimited and randomized within a certain text).

We want to specify again that the informants can include any word into any microtheme, which means that not only words but microthemes can be located in different parts of any text. Hence one can observe microtheme overlapping in some parts of text.

The obtained interpretations consist of semantic groups based on microthemes. The included words within one microtheme are linked with each other. Any word may be included in different microthemes. Word ability to be included into several microthemes allows us to speak about semantic connection intensity. The more frequently two words are included into one microtheme, the more semantic connection intensity they manifest. Semantic charting of the text shows semantic connection intensity of each word with other words of the text (based on all submitted interpretations) (see Table 1). In the given semantic chart the words from vertical and horizontal lines are taken from the text below. The figures at the crossing manifest the semantic connection intensity.

Лето умирает. Осень умирает. Зима – сама смерть. А весна постоянна. Она живет бесконечно в недрах вечно изменяющейся материи, только меняет свои формы. (В. Катаев)

Summer dies. Autumn dies. Winter is the death itself. And spring is constant. It lives infinitely in the womb of perpetually changing matter, it only changes its form. (V. Katayev)

It is evident that words can be connected either regularly or accidentally. The level of significance can be defined statistically. It should surpass the sum of mean value and mean square deviation of the semantic connection intensity for each word. For example, the word *лето* has the following semantic connection intensity with the other words: *лето* (0), *умирать* (5), *осень* (10), зима (8), *сама* (0), *смерть* (3), *а* (0), *весна* (5), *постоянна* (0), *она* (0), *жить* (2), *бесконечно* (2), *недра* (2), *вечно* (0), *изменяющийся* (0), *материя* (0), *только* (1), *менять* (1), *свои* (1), *форма* (0).

For instance, the cell of the column «жить» and the line «весна» gives the value «7». It means that these words were placed by the informants into one semantic group (microtheme) seven times (of possible 21 times according to the number of the informants).

The mean value of the semantic connection intensity of the word *лето* with other words equals to 2. The mean square deviation, equals to 2,88 (1-sigma). Hence semantic connection intensity can be regarded nonrandom as it exceeds 4,88. All the rest figures are accidental, that's why they are not considered.

Table 1. Fragment of the semantic charting Лето…

Word/word	лето	умирать	осень	зима	сама	смерть	а (but)	весна	постоянна	она (it)	жить	бесконечно	недра
лето (summer)	0	5	10	8	0	3	0	5	0	0	2	2	2
умирать (die)	5	1	7	4	0	7	0	0	0	1	2	1	1
осень (autumn)	10	7	1	9	1	5	1	3	0	1	0	1	0
зима (winter)	8	4	9	0	2	8	1	4	0	1	0	2	0
сама (itself)	0	0	0	2	0	3	6	3	2	8	0	2	4
смерть (the death)	3	7	4	8	2	0	1	0	0	1	2	3	1
а (but)	0	0	1	1	6	2	1	3	2	6	1	0	1
весна (spring)	5	0	4	4	3	0	2	0	8	3	7	4	2
постоянна (constant)	0	0	0	0	2	0	2	8	0	2	6	5	2
она (it)	0	1	1	2	8	1	6	4	2	0	2	1	3
жить (live)	2	1	0	0	0	2	2	7	6	2	0	6	3
бесконечно (infinitely)	2	2	1	2	2	3	0	5	5	1	6	1	1
недра (womb)	2	1	0	0	3	1	1	2	1	4	3	2	0

Semantic Graph of a Text

Due to the described technique the quantity of semantic connections of the words decreases substantially. It makes possible to use graphical means for the representation of semantic connections. To achieve it we place the words on the plane and show regular connections between them using connective lines (see Figure 2).

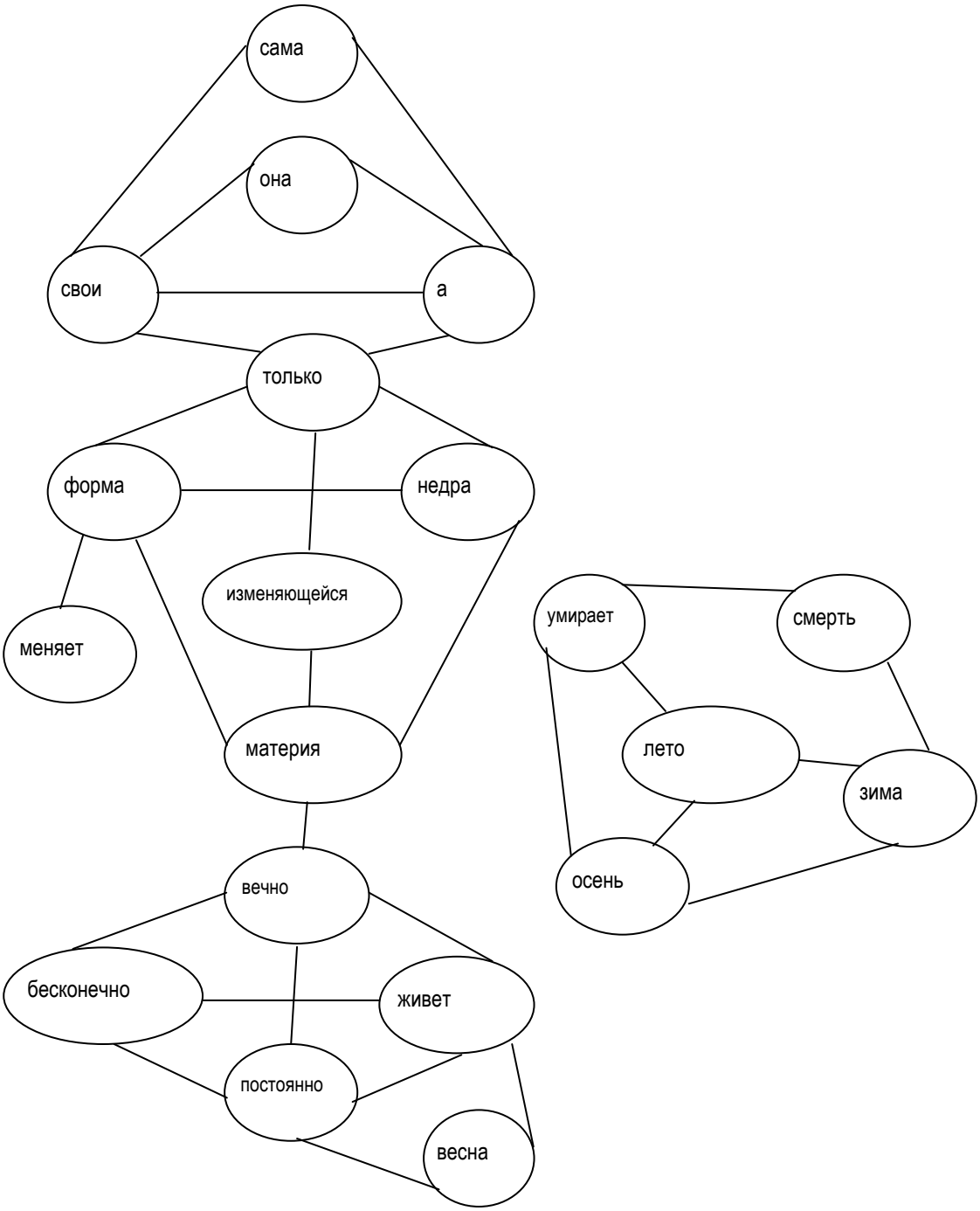

Figure 2. The semantic connections between the words of the text Лето…

In the Figure 2 we see that the semantic space structure consists of two substructures. The smaller one is the substructure, composed of the words *лето*, *осень*, *зима*, *смерть*, *умирать*, which can be conventionally denoted as the semantic field "death". This semantic field is formed only on the ground of contextual connections. The structure is homogeneous, no kernel element (the element having the greatest number of connections with other components of the same substructure) can be singled out in it. It is interesting to notice that this substructure is not connected with the dominant structure at all.

Evidently the "disconnected" semantic space provides the explanation of the fact, that there is considerable diversity in defining the theme of the text in the informants' interpretations (cf. some of the interpretations: *The text is about eternity of the spring. The text is about the spring being not only a season but something more. The text is about forms, which we, silly people, regard as different seasons. The text is about death of the seasons and eternal life of the spring. The text is about the infinity of the changeable creating spring .etc*).

As to the dominant substructure, it is discrete in the long run. So, we can single out the substructure formed by the words *вечно, бесконечно, постоянно, жить, весна*, which can be denoted as the semantic field "life". The substructure has the kernel element shown graphically. The kernel of the substructure, the word *вечно*, is the mediator between the given substructure and the main part of the dominant substructure. The dominant substructure also includes two smaller ones formed by the words 1) *материя, изменяющийся, недра, форма, только, менять* and 2) *только, свои, она, а, сама*. The difference between them is based on the functions performed by their components: those of the first mentioned substructure have a nominative function and components of the second substructure realize deictic and relative functions.

The tendency to be discrete is also seen in its syntactical organization. One can observe syntactical parallelism in the initial sentences which allows them to be more free compositionally than in case of chain connection. Having the adversative meaning the conjunction *а* also contributes to its discrete character as it divides the lingual matter into two parts. The conjunction *только* has the similar effect, although it has no adversative meaning, but it determines more precisely the notion *бесконечность* (*жизнь весны*) as infinity of its changing forms. And continuity is most clearly realized as the phenomenon, surmounting discontinuity which is observed most vividly in the syntagm *Она живет бесконечно в недрах вечно изменяющейся материи*. Here, on the sentence level, the text becomes syntactically deep, as opposed to the plane relations with coordinative connection only, realized in the first sentences (the tendency to discontinuity prevented them from expanding).

The kernel component in the substructure *материя, изменяющийся, недра, форма, только, менять* (field «материя») is the word *форма*, since it has the largest number of connections (6). However some of this word's connections are of little importance. These are the connections with the irredundant element *менять*, with the words *только*, *свои*. The latter components are the mediators of the substructure *только, свои, она, а, сама*, which is apparently important grammatically, rather than semantically. The most relevant element of the field «материя» ("matter") is the word *материя*, because it has the same function as the word *вечно*.

All this allows us to assume that the semantic space of the text is highly discrete. And surmounting this discontinuity allows the text to exist as a whole, being realized with the help of the mediating elements of the substructures "жизнь" ("life") and "материя" ("matter"). Thus the category of "wholeness" is realized by means of connections of kernel components *материя*, *вечно* and a number of adjacent components (*бесконечно, жить, постоянно, изменяющийся, форма, недра*).

Semantic Outline of a Text

With the semantic charting of the text and structural analysis of its semantic organization we can reconstruct the semantic outline of the text, which models the semantic connection intensity between words in the *progressive linear text development.* The following procedures have been implemented.

1. All the words (including recurrent ones) in their linear sequence in the text are placed in individual cells.

2. Only regular connections between words should be shown in the semantic charting of the text.

3. The regular semantic connections between text elements are put down in all cells situated between these words. If the words *лето "summer"* and *весна "spring"* have the intensity of connections which is equal to "5" (see Table 1) this meaning should be situated in the whole cell between these words. Thus we assume that *the semantic connection between two words in a text as a linear object inevitably covers all the space between the lexemes*. In case of recurrence of the words the procedure is to be done again.

4. Having completed this procedure the number of semantic connections are summed up. Summing is carried out within every individual column, having a word of the text in its top cell. Thus we can observe semantic connection density at any moment of the linear text development in the process of its perception (see the result of constructing semantic outline in Table 2).

Physical Outline of a Text

To scrutinize the *physical organization of a text* we used audio software which allowed us to analyse the prosodic aspects of the text. In the course of the research the informants were asked to recite the given text, as the recital assumes its interpretation. Thus the semantic interpretation is realized by sound intensity, pitch and duration of a sound. Only the first property was taken into account due to the systemic character of the speech intonation because of interdependency of all three above mentioned properties. Audio recording equipment allowed us to keep distance between mouth and microphone to obtain valuable material for the analysis. After the recording, the speech properties were analyzed by means of the specialized programs *Cool Edit Pro* and *Excel* according to the following algorithm: 1) the acoustic wave visually represented by Cool Edit Pro was segmented into parts, each of them being equal to the recorded words; 2) the points of maximum sound intensity (maximum amplitude) of the stressed syllables were determined within every segment, and the maximum values of the acoustic wave

intensity were put down in the Excel table. Thus every recital was processed and presented in the summary table, where every recital was represented as a dynamic outline of the text. Absolute values were converted into relative ones in order to compare different recitals with each other. Then the mean values were calculated from all relative values of sound intensity which we call the prosodic outline of the text.

Synchronization of Semantic and Physical Outlines of a Text

By comparing and contrasting synchronically semantic connection intensity and mean sound intensity of the obtained data we received the results given in Table 2.

Table 2. Semantic connection intensity and sound intensity values for the text *Лето…*

	Semantic connections	Sound intensity (mean)
Лето (summer)	0,33	0,96
умирает (dies)	0,60	0,75
Осень (autumn)	0,85	0,96
умирает. (dies)	0,82	0,73
Зима (winter)	0,77	0,76
сама (itself)	0,73	0,76
смерть.(death)	0,73	0,69
А (and)	0,60	0,71
Весна (spring)	0,70	0,84
постоянна.(is constant)	0,86	0,74
Она (it)	0,83	1,00
Живет (lives)	0,80	0,90
бесконечно (infinitely)	0,73	0,85
в недрах (in the womb of)	0,85	0,93
Вечно (perpetually)	0,90	0,72
Изменяющейся (changing)	0,93	0,84
материи,(matter)	1,00	0,67
Только (only)	0,93	0,93
Меняет (changing)	0,89	0,87
Свои (its)	0,79	0,72
Формы (forms)	0,42	0,63

For easier perception all the data from Table 2 are presented graphically (see Figure 3) The X-axis indicates the words in their linear sequence and their location in the text. The Y-axis indicates the intensity of semantic and prosodic processes (1 being the maximum value). Figure 3 shows the earlier stated differences and similarities in the above-mentioned aspects of the text.

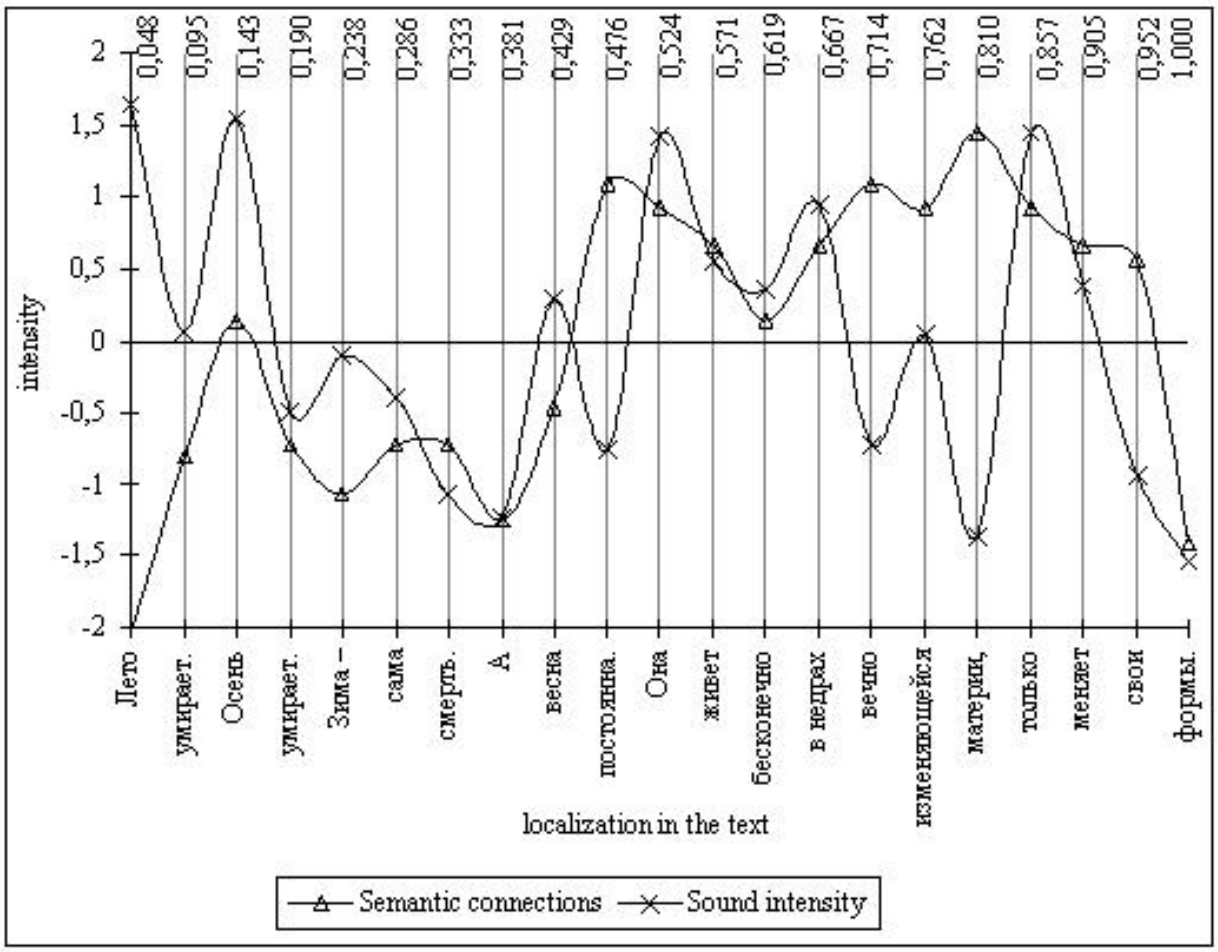

Figure 3. Intensity of semantic and prosodic processes in the linear development of the text.

We can see the three-part composition of the text in Figure 3. The three semantic maxima are situated between the semantic minima: *осень* – between *лето* and *а*, *постоянна* – between *а* and *бесконечно*, *материя* – between *бесконечно* and *формы*. According to the positional analysis of the text every text has the following position structure: the absolute beginning (AB) and the absolute end (AE), between which there is the harmonic centre of the text (HC, at the distance 0,618 from the beginning of the text), the harmonic centre of the initial zone (HCiz, at the distance 0,236 from the beginning of the text), the Setting (at the distance 0,146 from the beginning of the text), the absolutely weak positions (AWP_1, AWP_2, at the distance 0,236 to the right and to the left of the

HC) (see Figure 4). The mentioned above meanings are proportions of "golden section" which we use to define position structure of the text [Belousov, 2009].

We believe that a text is constructed according to the laws of harmony (proportions of "golden section") what affects text structure as a process and result of speech activity. This adjustment is known as text self-organization. One of self-organization markers is the creative attractor (the location in a text structure which has the greatest intensity of semantic process) of the text, understood as such extension in linear text space, in which self-organization processes are the most evident (explicit maximally). However the problem of the correspondence of the creative attractor to the units of language and semantic levels remains open. In Figure 3 we can see, that the HC of the text is the word *бесконечно*, HCiz – *зима*; AWP_1 is between *а* and *весна*, AWP_2 – between *только* и *меняет*. The position of the setting is between *осень* and *умирает*. The creative attractor occupies the interval between the HC and AWP_1 (*бесконечно в недрах вечно изменяющейся материи, только*). The above mentioned three-part composition of the text partially corresponds to its location structure: the initial zone of the text (from AB to AWP_1) is marked with the minimal number of semantic connections both to the left (the initial word) and to the right (the position between *а* and *весна*), i.e. it stands out of the whole structure; while the final zone of the text (from AWP_2 to AE) is not marked so clearly. There should be a considerable decrease of semantic connections intensity in absolutely weak positions, as their basic function is segmentation of the text structure. Something of this kind can be observed in the position AWP_1, while there are no distinctly seen mechanisms of discontinuity in AWP_2. Probably, the borders of the positions determined in the invariant must vary due to language substratum inertness: in our case the considerable drop in semantic connections intensity falls on the last word, and in this case AWP_2 should be between the words *свои* and *формы*. Regarding the harmonic center of the text (in Figure 3 HC is the word *бесконечно* "*infinitely*"), we see that semantic connection intensity in them appears to be of mean value. Harmonic centers had similar organization while we had analyzed other aspects of the text, its prosodic organization in particular.

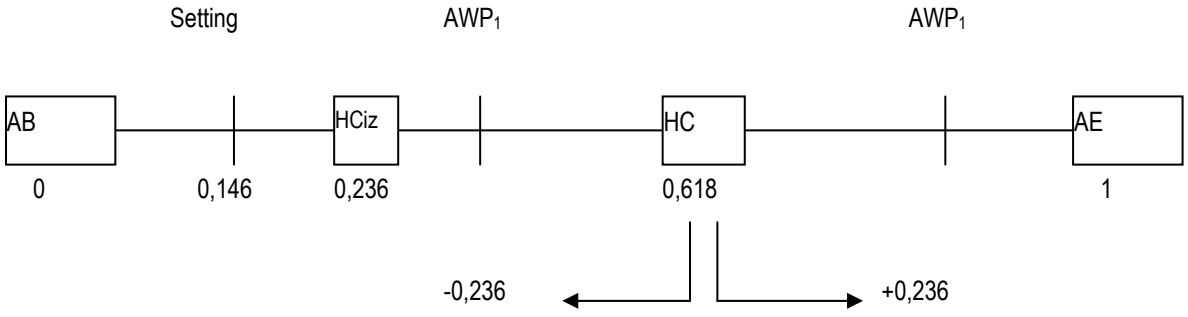

Figure 4. The position structure of the text

The creative attractor is situated in the interval with the maximum value of semantic connections intensity. In this case the attractor is in the interval where the intersection of all the semantic connections is the most evident, reflecting wholeness of this text in the utmost degree, what has already been discussed. However such processes are not characteristic for all texts.

As it is seen in Figure 3, the independent outlines begin to change interactively in the setting (0,146) of the text, and it can be interpreted as the attempt to coordinate their temporhythmical evolution parameters. We have already observed this function of the setting when we compared prosodic and emotional outlines of the text. It is interesting that the prosodic organization intensity in this interval is higher than the semantic one. In the HCiz the intensities of forming processes "equalize" for the first time and progress synchronically up to AWP_1 [cf. Belousov, 2008]. AWP_1, being the minimum point of the semantic and prosodic outlines, is the turning point of the development from which we can observe the synchronic growth of semantic and prosodic intensity. The prosodic processes dominate (their relative meanings exceed relative meanings of the semantic process (see Figure 3)) i.e. the semantic connection intensity tends to adjust to patterns of the physical aspect of the text development. Apparently such interaction can be called "suggestion". The coordinated intensity alteration of the two processes, registered in the interval around HC, indicates the harmonization of the processes, the stability point (the minimum) falling on HC. The area of the creative attractor (the maximum of semantic connections intensity) is the place of the most pronounced discrepancy of the semantic and prosodic outlines. The dominating role (their relative meanings the semantic process exceed those of the prosodic process) of the semantic component against the background of the prosodic phenomena shows the significance of cognitive mechanisms as opposed to emotive ones in this text self-organization. Strange as it may seem the area of the creative attractor comprises the words with abstract meaning: *бесконечно, недра, вечно, изменяющийся, материя*, which bring to the forefront rational-logical elements in the structuring of this text's wholeness. Besides, this interval has much more complicated syntactical structure than other intervals of the test. And finally in AWP_2 the intensity parameters of the both processes intersect and overlap each other to gradually dwindle synchronically up to the end of the text.

Conclusion

Thus the technique of semantic charting and the method of positional analysis allowed us to represent the successive-simultaneous semantic space of a text as its "semantic outline". Owing to the method of the prosodic analysis of a text, aimed at modeling its prosodic outline, there appears the possibility to analyze the cooperative interactions of these relatively independent text spaces.

By comparing and contrasting synchronically semantic connection intensity and mean sound intensity of the obtained data we received the results that allow us to be more specific in the discussion of the text structure as an evolving process. The search for explanatory tools of convergence, divergence, intersection, overlapping of various text structures is the key to understanding the complex material, ideal and social nature of text, its presentation as wholeness.

Bibliography

[Anochin, 1999] P.K. Anochin. Poleznyj rezul´tat kak organizujuščij faktor sistemy. In: Sinergetika i psichologija 2. Social´nye processy: 34-37. Moscow: JANUS-K2. 1999.

[Belousov, 2008] K.I. Belousov. Sinergetika teksta: ot strustury k forme. Moscow : Editorial URSS. 2008.

[Belousov, 2009] K.I. Belousov. Teorija i metodologija polistrukturnogo sinteza teksta. Moscow : Flinta: Nauka. 2009.

[Leont´ev, 2001] A.A. Leont´ev. Jazyk i rečevaja dejatel´nost´ v obščej i pedagogičeskoj psichologii. Moscow: Moskovskij psichologo-social´nyj institut. 2001.

Authors' Information

Konstantin I. Belousov - Orenburg State University, Department of Philology, Professor, Chief of Philological Modeling and Design Laboratory; P.O. Box 13, Prospekt Pobedy, Orenburg, 460018; e-mail: konstan-bel1515@yandex.ru

Major Fields of Scientific Research: Psycholinguistics and Psychosemantics, Semiotics, Semantics and Pragmatics, Cognitive Mechanisms for Text Processing

Tatyana N. Galinskaya - Orenburg State University, Department of Philology, Associate Professor; P.O. Box 15, Pereulok Yasnyi, Orenburg, 460036; e-mail: galinskaya@rambler.ru

Major Fields of Scientific Research: Psycholinguistics and Psychosemantics, Semiotics, Semantics and Pragmatics, Cognitive Mechanisms for Text Processing

APPLICATION OF MATHEMATICAL INDUCTION FOR INHERITANCE LAW INERPRETATIONS

Assen Tochev, Vassil Guliashki

***Abstract**: The purpose of this article is to obtain simple rule for applying the Inheritance law for the case of (own) brothers/sisters by birth, and/or brothers/sisters uterine or through father. Using the mathematical induction a result is obtained for n (own) brothers/sisters by birth and m brothers/sisters uterine or through father.*

***Keywords**: Inheritance law, mathematical induction.*

***ACM Classification Keywords**: A.0 General Literature - Conference proceedings; I. Computing methodologies, I.2. Artificial Intelligence, I.2.1. Applications and expert systems, Subject descriptor: Law; H. Information systems, H4. Information systems application, H.4.2. Types of systems, Subject descriptor: Decision support;*

Introduction

Let we consider the following part from the Bulgarian inheritance law (see [Law of inheritance, 1949] and [Tassev et al, 2009]):

Article 8.

(1) When the deceased has left only brothers and sisters, they inherit equal parts.

(2) When the deceased has left only brothers and sisters together with ascending in second or higher degree, the first obtain two thirds from the heritage and the last (the ascending) – one third.

(3) In the cases of foregoing clauses the brothers/sisters uterine and brothers/sisters through father obtain the half of the part inherited by the (own) brothers/sisters by birth.

(4) (New – State Newspaper, Nr. 60 since 1992) When the deceased has not left ascending in second or higher degree brothers and sisters or their descendents, inheritors are the relatives in collateral line till sixth degree inclusively. The closer by degree and the descending of one closer by degree excludes the more distant by degree relative.

Cases of heritage separation according the Bulgarian Inheritance law

1. Case (0,0) – **zero** (own) brothers/sisters by birth, and **zero** brothers/sisters uterine or through father.

PROBLEM: Let there are no brothers/sisters and/or brothers/sisters uterine or through father, who inherit. Who will obtain the heritage?

SOLUTION:

From Article 8 (4) can be concluded, that when the legator has not brothers and sisters, the relatives in collateral line till sixth degree inclusively inherit.

2. Case (0,1) – **zero** (own) brothers/sisters by birth, and **one** brother/sister uterine or through father.

PROBLEM: Let there is one brother/sister uterine or through father, who inherits. How big part from the heritage will inherit the brother/sister uterine or through father?

SOLUTION:

Let the brother/sister uterine or through father inherits 1/2 part from the heritage, as prescribed in Article 8 (3), i.e.:

Iteration 1

The rest of heritage is $Z_1=1/2$. The obtained heritage is $H_1=1/2$.

Iteration 2

The rest of heritage is divided by 2: $Z_2=1/4$. The obtained heritage is $H_2=1/2+1/4$.

Iteration 3

The rest of heritage is divided by 2: $Z_3=1/8$. The obtained heritage is $H_3=1/2+1/4+1/8$.

Iteration 4

The rest of heritage is divided by 2: $Z_4=1/16$. The obtained heritage is $H_4=1/2+1/4+1/8+1/16$.

...

Iteration *k*

The rest of heritage is divided by 2: $Z_k=1/2^k$. The obtained heritage is $H_k=1/2+1/4+\ldots+1/2^k$.

$$\lim_{k\to\infty} (1/2^k) = 0 \qquad (1)$$

Hence $1/2+1/4+1/8+1/16+\ldots+1/2^k = 1/2(1+1/2+1/4+\ldots+1/2^{k-1}) =$

$=1/2((2^{k-1}+2^{k-2}+\ldots+2^1+1)/ 2^{k-1}) =$

$=1/2(((2^{k-1}+2^{k-2}+\ldots+2^1+1).(2-1))/((2-1)2^{k-1})) =$

$=1/2((2^k-1)/2^{k-1}) =$ taking into account (1)

$=1/2(2^k/2^{k-1}) =$

$=1/2 . 2 =$

$=1$

Hence, when only one brother/sister uterine or through father inherits, he/she obtains the whole heritage.

This conclusion can be drawn also by another way:

Let we denote by X the part of heritage inherited by the (own) brothers/sisters by birth, and by Y the part of heritage inherited by the brothers/sisters uterine or through father. Then according Article 8 (1) and Article 8 (3) the following system of linear equations can be created:

X+Y=1

X=0

Y=1

3. Case (1,0) – **one** (own) brothers/sisters by birth, and **zero** brother/sister uterine or through father.

PROBLEM: Let there is one (own) brother/sister by birth and zero brothers/sisters uterine or through father, who inherit. How big part of the heritage will inherit the brother/sister by birth and the brothers/sisters uterine or through father?

SOLUTION:

Using X and Y as denoted in the previous case we create the following system of linear equations according Article 8 (1) and Article 8 (3):

X+Y=1

Y=0

X=1

Hence according Article 8 (1), when only one (own) brother/sister by birth inherits, he/she will obtain the whole heritage.

4. Case (1,1) – **one** (own) brothers/sisters by birth, and **one** brother/sister uterine or through father.

PROBLEM: Let there is one (own) brother/sister by birth and one brother/sister uterine or through father, who inherit. How big part of the heritage will inherit the brother/sister by birth and the brothers/sisters uterine or through father?

SOLUTION:

According Article 8 (1) and Article 8 (3) we create the following system of linear equations:

X+y = 1

y = (1/2).X

Hence:

X+(1/2).X = 1

(3/2).X = 1

X = 2/3

y = 1/3

It means that the brother/sister by birth obtains 2/3 from the heritage and the brother/sister uterine or through father – 1/3.

Now we will prove a simple rule for applying the Bulgarian Inheritance law in case of *n* (own) brothers/sisters by birth and *m* brothers/sisters uterine or through father by means of method of mathematical induction (see [Knuth, 1997]).

Common rule obtained trough mathematical induction

Let we have *n* (own) brothers/sisters by birth and *m* brothers/sisters uterine or through father.

We denote the part of heritage obtained by each brother/sister by birth by U. The part of heritage obtained by each brother/sister uterine or through father is denoted by V.

Statement 1: In the case of *n* brothers/sisters by birth each of them will inherit U=X/*n* part of heritage.

Statement 2: In the case of *m* brothers/sisters uterine, each of them will inherit V=Y/*m* part of heritage.

We will prove *Statement 1* by means of mathematical induction.

Proof of Statement 1:

1) Let we have *n* = 1 brother/sister by birth. Then according the *Case 3* he/she will obtain the whole heritage, i.e. the inherited part U is: U = X/1.

2) We assume that the Statement 1 is true for *n* = *k*.

We have X = *k*.U, or U = X/*k*.

3) We consider the case with *n* = *k*+1 brothers/sisters by birth.

Taking into account 2) X = *k*.U + U = U.(*k*+1).

Hence U = X/(*k*+1). ■

The Statement 2 can be proved by means of mathematical induction in a similar way.

Conclusion

Let the brothers/sisters by birth, and/or the brothers/sisters uterine or through father are *n* and *m* correspondingly. In all cases, when $n > 0$ and $m > 0$, the part of the heritage, obtained by the corresponding brothers/sisters is equal to X/*n* and У/*m* correspondingly.

In the same way similar rules for inheritance laws interpretations in other countries can be proved. The obtained rules can be used in artificial intelligence systems for expert consultations in the area of inheritance law.

Bibliography

[Law of inheritance, 1949] Law of inheritance, in effect since 30.04.1949, published in State Journal Nr. 22 as of January 29. 1949, corrected in State Journal Nr. 41 as of February 21. 1949, changed in State Journal Nr. 275 as of November 1950, changed in State Journal Nr. 41 as of May 28. 1985, changed in State Journal Nr. 60 as of July 24. 1992, changed in State Journal Nr. 21 as of March 12. 1996, changed in State Journal Nr. 104 as of December 6. 1996, completed in State Journal Nr. 117 as of December 10. 1997, completed in State Journal Nr. 96 as of November 5. 1999, changed in State Journal Nr. 34 as of April 25. 2000, changed in State Journal Nr. 59 as of July 20. 2007, changed in State Journal Nr. 47 as of June 23. 2009;

[Tassev et al, 2009] Tassev Ch., G. Petkanov, S. Tassev, Bulgarian Inheritance Law, Ninth overworked issue, Ciela, 2009.

[Knuth, 1997] Knuth, Donald E., The Art of Computer Programming, Volume 1: Fundamental Algorithms (3rd ed.). Addison-Wesley, 1997, ISBN 0-201-89683-4. (Section 1.2.1: Mathematical Induction, pp. 11–21.)

Authors' Information

Assen Tochev *– Institute of mathematics and informatics - BAS, "Acad. G. Bonchev" Str. Bl. 8, Sofia-1113, Bulgaria; e-mail: tochevassen@yahoo.com*

Major Fields of Scientific Research: Artificial intelligence systems, Legal informatics, Decision and information systems

Vassil Guliashki *– Doc. Ph.D., Institute of Information and Communication Technologies - BAS,*

"Acad. G. Bonchev" Str. Bl. 2, Sofia-1113, Bulgaria; e-mail: vggul@yahoo.com

Major Fields of Scientific Research: Discrete optimization, Evolutionary algorithms, Global heuristic strategies, Multiple objective programming, Decision support systems

OPTIMISATION OF ROUTE-PLANNING UNDER INDEFINITE RISK CONDITIONS

Kuzemin Oleksandr, Berezhnoy Sergey, Dayub Yasir

***Abstract**: This paper describes an algorithm of data transformation with a view to provide support for the decision maker. The aim of the paper is to develop a multi-purpose algorithm of building sets of optimal routes, taking into consideration most of the real factors that provoke risks. A simple and effective method of multicriteria optimization was proposed and developed.*

***Keywords**: emergency situations, microsituations, road conditions, weather conditions, objects of high danger, multicriteria optimisation.*

***ACM Classification Keywords:** H.1 Models and Principles – General*

Introduction

Nowadays vehicle route management in indefinite emergency situations uprising tends to be even more important. This problem includes modeling of transportations, decision-making optimization, and development of informational environment for decision-making support in field of high caution cargo transportation. The most of these problems is still solved by people, using their experience and intuition. Unfortunately, human mistakes take place and can dramatically influence the situation, especially in case of high caution cargo transportation.

Human factor is one of the most important and destructive risk-producing ones. Humans are strongly influenced by their emotions, their health state, their mood and other things, that actually, should not affect decision-making. That's why decision-making should be supported by mathematical models and methods that negate these problems as widely as possible.

Organization of safe routes for high caution cargo in case of emerging of an exceptional situation is one of the most important problems of traffic management, which requires updating and upgrading of approaches and decision methods used for this problem and usage of newest inventions in information technology.

The most frequently used method is still single-criteria optimization. Usually time or distance is optimized in order to meet client requirements. These methods do not need any information, concerning weather or road state or any other very important factors. So these methods simply ignore the most risk-emerging events. Of course this leads to great risks and therefore losses.

One-criteria optimization do not deal with risk, that is its major and fatal disadvantage. Modern transport can already provide speed and cheapness, but it is still not protected from risk.

Route management now needs a simple and effective tool for risk optimization. From single-criteria optimization we move to multi-criteria optimization, which is the key to accurate and effective way to avoid major risks, summoned by weather, catastrophes and other negative events.

The goal of this work is to provide the tool, which would possess the following features: computational optimality, educability, high adaptability, human factor taken into consideration, both changeable and unchangeable factors taken into consideration, ability to process data of different types.

We have developed this tool, it is easily adaptable for any kind of transport network. This work widely uses a term of "microsituation", which means a set of qualitative and quantative rates (which are characteristics of a microsituation) and a territory or a part of road, on which these rates can be considered as constant. Any change of parameters generates a new microsituation. So, they can evolve into each other under influence of some momentary events or conditions. These evolutions can be reduced to scenarios, because mainly, similar microsituations under similar conditions evolve into similar new microsituations.

Any road between two cities can be presented as a chain of different microsituations. Each microsituation carries its own risk. We have developed a way of aggregating these microsituations into one complex rate, basing on which it is possible to perform optimization.

We shall show the principle of our development on a simple problem.

The essence of work

The proposed method:

A system is presented in a form of an oriented graph, imitating a traffic network, a vehicle, moving in this network, weather conditions, emergency situations and objects of high danger. Set of possible decisions contains all possible edges. A set of scenarios S consists of all possible finite sets of sequential microsituations, where the starting microsituation is the current situation, surrounding the vehicle. We introduce a set of extreme situations and three metrics, which will be described a bit later. The most important point here is, that we use metrical distance between the examined microsituation and the extreme set to optimize risk. In some sense, norm, derived from all of three metrics is a kind of risk-measure. An extreme microsituation is such a situation, which has its risk level higher than allowed. [Кузёмин 2006].

Every microsituation x has a corresponding finite fuzzy set of microsituations, which gives a finite number of possible system development scenarios. A decision is a chain of fuzzy sets of microsituations. To simplify, we

suppose, that we can fold a fuzzy set into one non-fuzzy microsituation, using their belong function. So, than, decision is a microsituational chain.

So, the problem looks like:

$$ch^* = arg\left(\min_{ch \in CH} \int_{x \in ch} |x| \right),$$

where CH is a set of all allowed chains.

Scenarios and transition probabilities can be defined either by an expert estimate, or using statistical data about early routes and microsituational sequences.

Weight vector is built using educational algorithms. Constraints are set by expert estimates.

Now we identify criterion, used by decision-maker, while searching for optimal decision. We have three main criterion: extreme situation (of natural or technical origin) risk, traffic accident risk (involving our vehicle only), robbery/stealing risk. All three criterion depend on system parameters and can be computed for every microsituation, and respectively for every decision. In addition to system parameters we use a priori probabilities of all three risk types in form of either statistical estimates or interval estimates. [Кузёмин 2007]

Now we shall clarify criterion vector more thoroughly. Let f_A, f_E, f_R be criterion functions for an accident, an extreme situation and a robbery. A priori probabilities are Q_A, Q_E, Q_R respectively. Vector of weather variables $\overline{Wt}$, a cumulative visit coefficient K_R, time t, passed distance s. In addition we introduce an extreme situation object. This type of object is a bounded area, in which it affects microsituation parameters accorfing to the expert-defined function depending on object's special parameters and microsituation parameters. Let it be E.

Here we have defined all the factors, which form preference relation for decision set in every microsituation.

Every microsituation carries a set of parameters, significancy of which depends on a certain scenario. Here they are: roadbed quality, road profile complexity, presence of high danger objects, astronomical time, weather conditions, surrounding relieve, crime rate, traffic load, visibility, roadbed state. [Hamdy 2007] These parameters are variously connected with each other and are used for computing criterion values.

Being in microsituation x we solve the following problem:

$$\mathrm{ch}^* = \arg\left(\min_{\mathrm{ch} \in \mathrm{CH}} \int_{x \in \mathrm{ch}} \alpha_A f_A^r \left(\overline{\beta_A} \times \left(Q_A, E, t, s, \overline{Wt} \right) \right) + \alpha_E f_E^r \left(\overline{\beta_E} \times \left(\overline{Wt}, E, Q_E \right) \right) + \alpha_R f_R^r \left(\overline{\beta_R} \times \left(Q_R, K_R \right) \right) \right),$$

where CH is a set of chains, beginning in x. Here $\overline{\beta_A}, \overline{\beta_E}, \overline{\beta_R}$ - factor weight vectors.

After decomposing factors down to parameters we obtain the following problem:

$$\text{ch}^* = \arg\left(\min_{\text{ch}\in\text{CH}} \int_{x\in\text{ch}} \overline{\alpha}(f_A^r\left(\overline{\beta_A}\times\left(Q_A,E,OHD,t,s,\overline{Wt},Rc,Rq\right)\right),\right.$$
$$\left.,f_E^r\left(\overline{\beta_E}\times\left(\overline{Wt},E,Q_E,OHD\right)\right),f_R^r\left(\overline{\beta_R}\times\left(Q_R,H,CR\right)\right)\right)$$

Here Rc, Rq - road profile complexity and roadbed quality respectively. OHD – objects of high danger, H – pass history, CR – crime rate.

Now, let's introduce metrics, that we will use during optimization.

Low-level metrics is a simple weighted Euclidean metrics of order 2, for the set of vectors $\left(Rc,Rq,OHD,H,CR,Q_A,Q_E,Q_R\right)$ with weight vector α_L.

Middle-level metrics is a weighted Euclidean metrics of order 2, for the set of vectors

$\left(Rc,Rq,OHD,H,CR,Q_A,Q_E,Q_R,\overline{Wt},K_R,E\right)$ with weight vector α_M.

High-level metrics is defined using norm

$$\left|x\right| = \overline{\alpha}(f_A^r\left(\overline{\beta_A}\times\left(Q_A,E,OHD,t,s,\overline{Wt},Rc,Rq\right)\right),f_E^r\left(\overline{\beta_E}\times\left(\overline{Wt},E,Q_E,OHD\right)\right),f_R^r\left(\overline{\beta_R}\times\left(Q_R,H,CR\right)\right)$$

Optimization Algorithm:

1. Defining the set of decisions.
2. Defining the extreme set.
3. For each decision we obtain a low-level distance between its results and extreme set.
4. Decisions, which are too close to the extreme set are dropped.
5. Using middle-level metrics we drop some more decisions.
6. Using norm, derived from middle-level metrics, we obtain speed limit for each of the microsituations.
7. Using speed limit we simulate vehicle movement and microsituations' evolution and compute distance, passed by the vehicle in each decision.
8. Decisions, which do not meet the constraints, imposed on daily distance and schedule, are discarded.

9. Using time, driver's state, speed limit and first-level risk function we obtain high-level distance between decision and extreme set and high-level norm for each of the microsituations.

10. Using high-level norm we obtain an integral estimate for the risk.

11. We choose decision, which has the lower high-level norm within given set.

Sphere of application:

1. Informational intellectual systems for emergency situation control. Carriers will be able to change their route in case of emergency immediately, not waiting for the traffic controller order. If any of current expeditions faces an extreme situation, the route changes can be applied to all following expeditions, wherever that expeditions are at the moment. This will reduce the risk of more than one expedition encountering extreme situation.

2. Notification systems. Having a big enough database of emergency situations and simply route-passes will allow to forecast situation development for definite time horizon. After reaching some size the database will allow to minimize risks at the stage of their uprising. The model will gain ability to take into consideration periodic weather changes, seasonal winds, snow melting and so on.

3. Traffic scheduling systems. In case an expedition consists of more than one vehicle, we will be able to diversify risks by using different routes for different parts of one expedition. This will increase the possibility of successful delivery. Moreover, it will make the database reach forecast-size even faster.

4. A single unified route-pass-database. Emergency situation processing method will lead to a database, which will be able to be used for commercial applications and platforms. Development of a common emergency and route-pass description protocol will improve personnel management and vehicle management beginning from vacation scheduling and crew planning and ending with scheduling inspections and vehicle choice for any specific run.

5. Personnel education and support in case of emergency. Using an automated decision-making system will decrease human mistake probability. Any human-specific mistakes will be avoided and risks therefore decreased.

Municipal departments will use obtained information for preventive measures planning in regions of high-level risk. This also will help road building planning. Road creation and modification with a view to a priori emergency risk will make roads even safer.

Emergency department will get an effective instrument of rescue operations planning and their own infrastructure impact.

Scientific novelty:

Usage of multicreteria optimization instead of one-criterion and muti-stage optimization. While estimating risk levels different factors and characteristics of microsituations are taken into consideration, for example: roadbed quality, road profile complexity, visibility, astronomical time, weather conditions and so on. Optimization is performed in three stages to decrease computational charge. Model of estimation and algorithm of optimization are developed as sets of connected separate blocks; each of them has an ability to be educated and modified. Estimation method can also be obtained from gathering and processing expert estimations. The model can be clarified and completed with any amount of blocks, built within the same principle. Any block can be independently modified in case decision-maker's preferences change. Using simulation in target function computing: for more adequate results we simulate expedition movement using speed limit modified by risk level and human fatigue. After that we perform additional optimization with simulation results taken into consideration.

Practical value: The developed method can be widely spread, in cargo, cash and passenger transportation with some transportation type adjustment. The model allows changing route and minimizing risks in immediate real time. Carrier will decrease insurance charges, use its human and technical resources more efficiently. A similar model can be embedded as a social service. This service will notify about any non-standard situations on the current route and suppose alternative actions in real time. Combined with GPS-navigation this service will help avoiding overloads of certain road sectors, kilometer-long traffic jams behind traffic accidents on highways, it will help rescuers to react immediately. Every driver and every car will act as a sensor, transmitting data about surrounding microsituation. Resection of emergency situations' effects will be speeded up due to decrease of traffic. Finally, if this model is widely embedded, having enough computational capacity, it will principally change methods of route-planning.

Figure 1 shows examples of testing the algorithm on a simple network with edges of close length and characteristics. This was done so to illustrate ability to avoid extreme situations by the algorithm.

Conclusion

In this work we have created and developed a method of multicriteria optimization route planning. The main advantage of the method is it's high adaptability, educational ability and low computational charges. It overbears methods of one-criterion optimization due to the fact that it allows both minimizing risks and optimization not only by distance, but also by time. We have developed the described algorithm using Wolfram Mathematica for Students. The model successfully performed on simple networks with randomly generated disturbances (extreme situations) of unified structure.

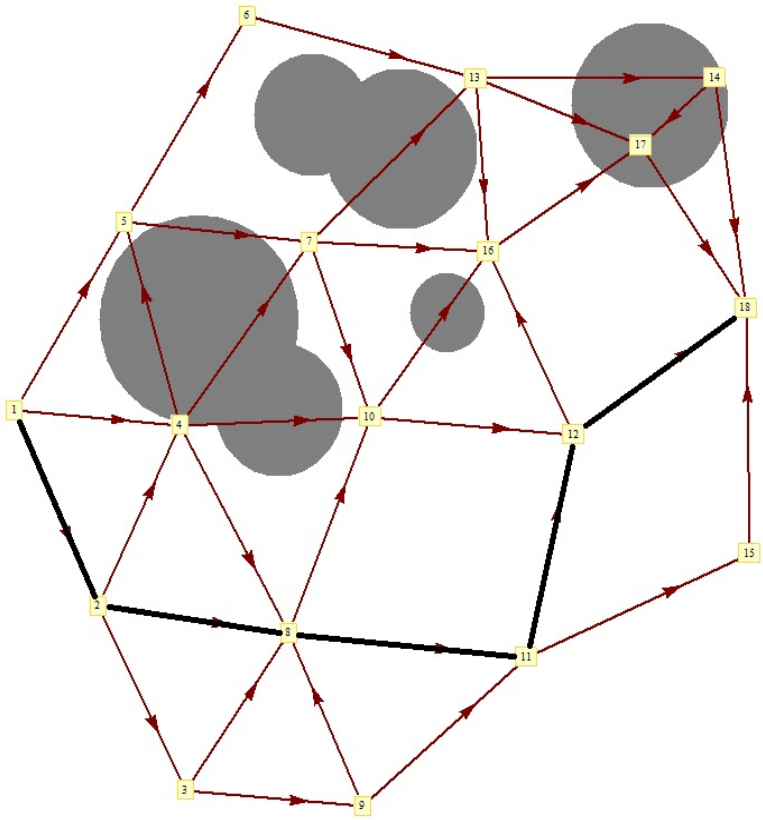

Figure 1 Example of a Test Map - randomly generated extreme situations with solution track.

Acnowledgements

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Кузёмин 2006] Кузёмин А.Я., Фастова Д.В., Дяченко О.Н. Геоинформационная система для классификации и прогнозирования лавинной опасности, АСУ и приборы автоматики. 2006.

[Кузёмин 2007] Кузёмин А.Я, Левыкин В.М. Разработка инструментальных средств обеспечения принятия решений для предупреждения и управления в чрезвычайных природных ситуациях, Автоматизированные системы управления и приборы автоматики. 2007.

[Кузёмин 2007] Кузёмин А.Я. Обобщённая модель и прогнозирование чрезвычайных природных ситуаций, Автоматизированные системы управления и приборы автоматики:, 2007.

[Hamdy 2007] Hamdy A. Taha. Operations Research: An introduction. 2007.

[Pidd 2003] Pidd M., Tools for Thinking: Modelling in Management Science, 2003.

[Kuzemin 2005] Kuzemin A., Sorochan M.,Yanchevkiy I., Torojev A. The use of situation representation when searching for solutions in computer aided design systems International Journal, Information Theories & Applications, 2005.

[Kuzemin 2005] Kuzemin A.., Fastova D., Yanchevsky I. Methods of adaptive extraction and analysis of knowledge of knowledge-base construction and fast decision making, International Journal on Information Theories & Applications, 2005.

Authors' Information

Oleksandr Kuzomin

Chief of Innovation Marketing Department

Professor of Information Science

14, Lenin Ave., 61166, Kharkiv, UKRAINE

Tel/fax: +38(057)7021515

mailto:kuzy@kture.kharkov.ua

Berezhnoy Sergey – Kharkiv National University of Radioelectronics; Kharkiv, Ukraine;

e-mail: serg.berezhnoy@gmail.com

tel.: +380 99 03 09 03 2

Major Fields of Scientific Research: Risk Management, Market Risk Optimization, Operational Risks Management.

Dayub Yasir. - postgraduate student; Kharkiv National University of Radioelectronics; Kharkiv, Ukraine;

14, Lenin Ave., 61166, Kharkiv, UKRAINE

Tel/fax: +38(057)7021515

mailto:kuzy@kture.kharkov.ua

TABLE OF CONTENT