THIRD INTERNATIONAL CONFERENCE ON

QUANTITATIVE LINGUISTICS

AUGUST 26-29, 1997, HELSINKI, FINLAND

RESEARCH INSTITUTE FOR THE LANGUAGES OF FINLAND

EDITORS

Pauli Saukkonen

Helena Suni

Helsinki, Monila 1997

THIRD INTERNATIONAL CONFERENCE ON QUANTITATIVE LINGUISTICS

Porthania, Yliopistonkatu 3, Helsinki

PROGRAMME

Monday, August 25

17.00–18.00 REGISTRATION

Tuesday, August 26

9.00–10.00 REGISTRATION

10.00–10.30 OPENING WORDS

For the abstract, see page mentioned after each name

PLENUM I (Chairman: Reinhard Köhler)

10.30–11.00 Marcus, Solomon: How quantitative is quantitative linguistics? ***10***

11.00–11.30 Tuldava, Juhan: Investigating causal relations in language with the help of path analysis ***15***

11.30–12.00 Dolinskij, Vladimir A. & Drogalina, Zhanna A.: V. V. Nalimov's probabilistic model of language and conscience ***21***

12.00–12.30 Baayen, R. Harald & Tweedie, Fiona J.: The parameters of LNRE models as sample-size invariant: problems and opportunities ***29***

12.30–14.00 LUNCH BREAK

SECTION 1

Phonetics, Phonology (Chairman: Johannes Gordesch)

14.00–14.30 Nettle, Daniel: Coevolution of phonology and the lexicon in twelve languages of West Africa ***36***

14.30–15.00 Matsunami, Shinya: On pitch extraction ***42***

15.00–15.30 Iivonen, Antti: A computer representation of articulatory gestures related to word phonotactics and speech sound classes ***48***

15.30–16.00 COFFEE BREAK

16.00–16.30 Tambovtsev, Yuri: The distances between the Finno-Ugric languages derived on the basis of distribution of consonantal group patterns in their speech chains ***51***

SECTION 2

Semantics (Chairman: Sheila Embleton)

14.00–14.30 Saukkonen, Pauli: A semantic typology of Finnish verbs: Statistics for role frames ***54***

14.30–15.00 Ljalkova, I.: Groups of English verbal synonyms ***56***

15.00–15.30 Ogoui, Alexander: Bestimmung der Antonymiebeziehung zwischen den polysemen deutschen Wörtern ***59***

15.30–16.00 COFFEE BREAK

16.00–16.30 Krott, Andrea: The influence of morph-polysemy on morph-frequency ***61***

16.30–17.00 Dolinskij, Vladimir A.: Russian words "russkij" and "sovetskij": Semantics of noun as an object of quantitative analysis ***68***

17.15 RECEPTION by the University of Helsinki

Wednesday, August 27

PLENUM II (Chairman: Pauli Saukkonen)

9.00–9.30 Králík, Jan: On the probability of probabilities. Some notes to stochastic approach to the Zipf formula ***77***

9.30–10.00 Prün, Claudia: G.K.Zipf's conception of language as an early prototype of synergetic linguistics ***83***

10.00–10.30 Skousen, Royal: Natural statistics in language modelling ***90***

10.30–11.00 COFFEE BREAK

11.00–11.30 Köhler, Reinhard: Syntactic structures: properties and interrelations ***92***

SECTION 1

Grammar (Chairman: Harald Baayen)

11.30–12.00 Hug, Marc: The French demonstrative particles -ci and -là: linguistic intuition and statistical facts ***94***

12.00–12.30 Hoffmann, Christiane: Word order and the principle of "Early immediate constituents" ***96***

12.30–14.00 LUNCH BREAK

14.00–14.30 Juola, Patrick: Measuring linguistic complexity: The morphological tier ***98***

14.30–15.00 Klavina, Sarma: Das Inventar der morphemischen Worttypen im Lettischen –

15.00–15.30 Budzhak-Jones, Svitlana: Multivariate rule analysis of case assignement in bilingual discourse ***100***

15.30–16.00 COFFEE BREAK

Text analysis (Chairman: Juhan Tuldava)

16.00–16.30 Cornelis, Louise & Bergh, Huub.v.d.: Dependent observations in text linguistics: the occurrence of the passive in Dutch ***108***

16.30–17.00 Tolcsvai Nagy, Gábor: Quantity and style from a cognitive point of view ***113***

17.00–17.30 Larouk, Omar: Logico-semantics and statistics applied to textual data in information retrieval ***118***

17.30–18.00 Pawlowski, Adam: Language in the line vs. language in the mass. On the efficiency of sequential modelling in the analysis of rhythm ***129***

SECTION 2

11.30–16.00 Round table discussion (Chairman: Sheila Embleton) ***136***

Regularities in natural language dynamics

Polikarpov, Anatoliy A.: Major tendencies in micro- and macro-dynamics of natural language lexical system ***138***

Forms (Chairman: Ludmila Uhlířová)

16.00–16.30 Kageura, Kyo: Type-based and token-based learning of Kanji morphemes ***146***

16.30–17.00 Meyer, Peter: Relating word length to morphemic structure: A morphologically motivated class of discrete probability distributions ***152***

17.00–17.30 Leopold, Edda: Frequency spectra within word length classes ***156***

18.15–19.45 SIGHTSEEING

19.45–22.00 RECEPTION by the Research Institute for the Languages of Finland

Thursday, August 28

PLENUM III (Chairman: Jan Králík)

9.00–9.30 Uhlířová, Ludmila: Linguists vs. the public: An electronic database of letters to the Language Service as a source of sociolinguistic information ***158***

9.30–10.00 Embleton, Sheila & Wheeler, Eric S.: Multidimensional scaling methods applied to a computerized dialect atlas of Finnish ***161***

10.00–10.30 Wimmer, Gejza & Altmann, Gabriel: An explorative method concerning word classes ***166***

10.30–11.00 Mukherjee, Sibasis: Structure of language – a quantitative approach ***167***

11.00–11.30 COFFEE BREAK

Forms (Chairman: Kimmo Koskenniemi)

11.30–12.00 Best, Karl-Heinz: Zum Stand der Untersuchungen zu Wort- und Satzlängen ***172***

12.00–12.30 O'Boyle, P. & Ming, J. & Owens, M. & Smith, F.J.: Adaptive parameter training in an interpolated N-gram language model ***177***

13.00–16.00 CRUISE

16.30–17.30 IQLA business meeting

NIGHT OF ARTS

Friday, August 29

SECTION 1

Lexicon (Chairman: Anatoliy Polikarpov)

9.00–9.30 Forster, Peter: Network analysis of vocabulary lists ***184***

9.30–10.00 Chizhakovski, Valentin A. & Popescu, Anatol. N.: Rule and network oriented approach to the semiotic model of the linguistic sign ***187***

10.00–10.30 Sanada-Yogo, Haruko: Analysis of Japanese vocabulary by the theory of synergetic linguistics ***193***

10.30–11.00 COFFEE BREAK

Lexicon (Chairman: Marc Hug)

11.00–11.30 Polikarpov, Anatoliy A.: Semasiological and word-formational processes in natural language lexical evolution ***194***

11.30–12.00 Shelov, S.D.: On measuring "termness": A quantitative approach to "term–nonterm" controversy ***199***

12.00–12.30 Devos, Filip & Maesfranckx, Patricia & De Tré, Guy: On granularity in the interpretation of around in approximative lexical time indicators ***203***

12.30–14.00 LUNCH BREAK

14.00–14.30 Komarnytska, Larisa: Factors determining phonetic motivation of the words: An experiment in phonetic symbolism ***210***

14.30–15.00 Krylov, Yu. K.: On the nature of Köhler's Effect ***214***

15.00–15.30 Krylov, Yu. K.: Three laws of fiction prosaic texts organization ***216***

15.30–16.00 COFFEE BREAK

16.00–16.30 Pokrovskaya, Elena A.: Database of Russian synonyms and its quantitative systemic analysis ***218***

16.30–17.00 Chebanov, Sergei V.: Text as real population in A. A. Chuprov sense ***220***

17.00–17.30 Seppänen, Jouko: Model Thinking as the Qualitative Foundation of Linguistics ***223***

SECTION 2

Psycholinguistics, Language acquisition

11.00–11.30 Choudhry, Amitav: The chi-square test and its significance in studying stability in response patterns ***226***

11.30–12.00 Kunsmann, Peter & Gordesch, Johannes & Dretzke, Burkhard: Native Speakers' reaction to modern English usage ***228***

12.00–12.30 Dolinskij, Vladimir A. & Rudakov, Sergey S.: Associative linguistic experiment and elaboration of methods of computer diagnostics for human's inborn hereditary syndromes ***233***

12.30–14.00 LUNCH BREAK

14.00–14.30 Holleman, Bregje C.: Is forbidding not allowing? And why not? A meta-analysis ***239***

14.30–15.00 Gieseking, Kathrin: Evaluating a frequency-based principle of human sentence processing ***243***

15.00–15.30 Wakabayashi, Setsuko: Quantitative analysis of different processing patterns in listening comprehension by L2 listeners ***249***

15.30–16.00 COFFEE BREAK

16.00–16.30 Niemi, Jussi: Quantitative aspects of Finnish Wernicke speakers' narratives ***255***

16.30–17.00 Otlygin, Vladimir: Trajectories of fractal attractors in preferability situations ***257***

17.00–17.30 Ranjita Nayak: Linguistic annotation of text corpora: a probabilistic approach –

17.30– CLOSING WORDS of the new President of the IQLA Council

18.00–19.00 RECEPTION by the City of Helsinki (Vanha Raatihuone)

EVENING IN A RESTAURANT

PROGRAMME

Gabriel Altmann, chairman

Reinhard Köhler

ORGANISATION

Pauli Saukkonen, chairman

Raimo Jussila

Fred Karlsson

Kimmo Koskenniemi

Ritva Liisa Pitkänen

Tuomo Tuomi

SECRETARIES

Pirkko Iivanainen

Marja Noronen

PLENUM I

HOW QUANTITATIVE IS QUANTITATIVE LINGUISTICS ?

Solomon Marcus

"Quantitative" is opposed to "qualitative" and, as an attribute of the reality investigated in science, is of two types: deterministic and probabilistic.Let us consider this attribute in respect to linguistics.

When was the syntagm "Quantitative Linguistics" used for the first time and in what circumstances ? We cannot give a precise answer, but we remember that the journal "Prague Studies in Mathematical Linguistics", started in 1965, had two sections: Quantitative Linguistics and Algebraic Linguistics. However, the former section was (and is) almost exclusively devoted to probabilistic, statistical and information aspects of language, while the latter wa (and is) predominantly dedicated to logical, algebraic and set-theoretic models of language. The syntagm "algebraic linguistics" was introduced by Y. Bar-Hillel,in order to delimitate that part of mathematical linguistics which is rather concerned with qualitative (i.e., logical, algebraic, topological etc.) than quantitative-statistical aspects of language. This happened in the fifties, the period of emergence of qualitative mathematical linguistics (that will be described later), in contrast with the long tradition of the probabilistic-statistical approach to language (frequency dictionaries, statistical stylistics etc.).

But things were not at all clear. For instance, in the same period, the Swedish journal "Statistical Methods in Linguistics" (SMIL) included all types of articles of mathematical linguistics, many of them completely away not only from probability and statistics but also from any quantitative approach, be it deterministic or probabilistic; see, in this respect, the articles of S. Kanger 1962 and S. Marcus 1965 concerned with a model of the phoneme. This situation can be understood in respect to a double assimilation: "mathematical" was assimilated with "quantitative", in view of a long tradition, going back to the past century, while "quantitative" was confused with "statistical", because the most elementary approach to quantity is the operation of counting, leading directly to the simplest idea of statistics.

When, where and in what circumstances was coined the label "mathematical linguistics" ? Most American authors avoided id, despite the fact that they did not ignore the strong link between mathematics and linguistics. A.G. Oettinger publishes in 1957 the article "Linguistics and Mathematics", while Roman Jakobson is in 1961 the editor of "The Structure of Language and its Mathematical Aspects". Only later on, in 1968, R. Abernaty publishes the article "Mathematical Linguistics" (but concerning the work done in Soviet Union) and R. Wall publishes in 1972 the book "Introduction to Mathematical Linguistics". In Europe, already in 1959 the soviet philosopher publishes an article "The signi-

ficance of mathematical linguistics", while P. Braffort publishes in 1960, in Belgium, the report "Elements de linguistique mathematique"; the same year, in Soviet Union, a decision of the Ministry of Higher Education is called: "On the formation of specialists and the development of scientific research in the field of mathematical and structural linguistics", while S. Marcus publishes (in a linguistic journal) the article "Some significations of mathematical linguistics" and three years later, in 1963, the book "Mathematical Linguistics". At the Ninth International Congress of Linguists (Cambridge, Mass.,1962) H. Spang-Hansen entitled his contribution: "Mathematical Linguistics - a trend in name or in fact ?"

The way towards mathematical linguistics was prepared from many directions: by linguists aiming a structural approach to language, for instance the axiomatic approach proposed by L. Bloomfield 1926, B. Bloch 1948, W. Harwood 1955,

J.H. Greenberg 1959 or linguistics seen as an algebra, by L. Hjelmslev 1943; by logicians bridging gradually logic and linguistics, such as Ajdukiewicz 1935, Y. Bar-Hillel 1950,1953,1954, J. Lambek 1958,1959,1961, by cyberneticians
such as V. Belevitch 1955, 1956, aiming to bridge machine language and human language.

In respect to the probabilistic-statistical source of mathematical linguistics, ignoring the events before 1940, we have to quote G.U.Yule 1944, P. Guiraud 1951,1959 (both with large impact among linguists and literary scientists) G.Herdan 1956,1960,1962. However, many studies of this period, claiming to use statistical tools,have a low degree of scientific accuracy and this situation determined Richard von Mises to joke saying that statistics is a form of lie. Many linguists became reluctant in respect to statistics (see the evolution of glotochronology). Only in a second step, a more rigorous approach to probabilistic aspects of language was obtained: J.P.Benzecri 1964, Ch. Muller 1968 and B. Brainerd 1975 are only some of them; the latter author is also the editor of a collective volume 1983, important for a basic reconsideration of the stochastic nature of phenomena in historical linguistics (this volume is number 18 in the prestigious series "Quantitative Linguistics", consacrating the respective label).

In order to understand the double face of the attribute "quantitative", as a symptom, some times, of a progressive step, other times as a conservative (if not regressive) one, let us look a little at the history of science.

Long time, and, some times, even in our days, the passage from observation and description to counting and measure was considered as an obligatory first step in a scientific (mathematical) approach. Particularly, mathematics was viewed as a science of quantitative and spatial aspects of reality. To a large extent, this picture of science and of mathematics was adequate for the Gali-

leo-Newtonian period and for its extension to the time of industrial and post-industrial revolution. But already in the XIXth century the discovery of non-Euclidean geometries, Galois' concept of a group, Felix Klein's Erlangen program and Poincare's approach to differential equations are major events showing a strong trend of mathematics from quantity and measure to quality and structure. A similar trend can be observed in chemistry (the discovery of the nature of chemical link and the elucidation of the phenomenon of isomerism), in physics and in biology. The XXth century, with the marginalization of matter by energy (Einstein) and, later, of energy by information, brings a fundamental switch to quality and structure. In this way, mathematics was faced with the need to enhance its stress on qualitative, formal aspects (topology, modern algebra, mathematical logic, graph theory etc.) and so, towards the middle of the XXth century, qualitative structural mathematics became predominant; the Bourbaki movement was a major aspect of this phenomenon.

Linguistics followed a similar orientation, but, in view of its late development,it has to cope concomitantly with these two opposite requirements:from description to measure and from quantity to quality. The former requirement started from linguistic data processing to process them by the tools of probability theory and mathematical statistics, while the latter requirement lead to the development of linguistic structuralism, as a preliminary step to qualitative mathematical modeling. So, these two steps, successive for mathematics, became rather simultaneous for linguistics.

The culminating moment of this process occurs towards the middle of the XXth century, with the emergence of the information paradigm; just within this framework appears mathematical linguistics and its adjacent fields (computational linguistics, language technology, linguistic engineering, theoretical linguistics, formal linguistics, applied linguistics etc.; the choice depends on the background and main interests of the respective researchers, scientific or engineering,linguistic or computational, theoretical or applied etc.).It is symptomatic in this respect that Chomsky's pioneering articles were published in IRE Transactions on Information Theory 1956 and Information and Control 1958,
1959, while the first Russian articles of mathematical linguistics were published in Biulleten Obiedinenia po problemam mashinogo perevoda (R.L.Dobrushin, V.A.Uspenskii 1957 and in Problemy Kibernetiki (O.S.Kulagina 1958). So, linguistics was bridging at that moment the new information fields called Information Theory, Cybernetics and Computer Science. The same thing can be said for the probabilistic approach to language; it acquires a new dimension within the framework of the information paradigm (see Shannon 1950, for the entropy of English), as it can be seen from L.Apostel-B.Mandelbrot-A.Morf 1957, L. Brillouin 1957, N.Chomsky-G.A.Miller 1958, O.S.Achmanova-I.A.Mel'chuk-E.Padu-

cheva-R.M.Frumkina 1961.

In order to capture the sense of the evolution of mathematical linguistics and, within this framework, the dynamics of interactions among quantitative and qualitative, let us sketch the successive waves of the information era (we will describe only the first three of them, leaving aside here the next three waves).

Information can be defined only in a negative way. Provisionally, it is what cannot be reduced to matter and energy. The first information wave, in the fourties, consists of the emergence of computer science, of cybernetics, of information theory, of coding theory (algebraic and probabilistic) theory, of molecular genetics and of a few more fields. This was just the historical and scientific context prepairing the development of mathematical, computational and quantitative linguistics, as recognized scientific fields. The dominant paradigms imposed by the first wave were "information", "computation" and "communication". Two facts are fundamental in this respect: a)computation is no longer, as in the past century, quantitative, i.e., dealing with numbers, it is qualitative, dealing with abstract symbols (in the tradition opened by A.M.Turing, in the thirties, prepared, in his turn, by Descartes, Leibniz and Boole); as a consequence, computer science is a qualitative field; b)the great merit of Shannon was the successful separation between information and meaning, leading to the possibility to measure the information. All further attempts(Carnap, Bar-Hillel, Hintikka etc.) to bridge information and meaning failed: a quantitative approach to meaning misses its semantic component and captures only the selective ones (of the type: "a statement says about empirical reality exactly what it prohibits about it"). The work of A.Moles and M.Bense is significant in this respect.

The second wave of information era, in the fifties, includes the emergence of Chomsky's generative linguistics, automata theory, cognitive science, artificial intelligence. These new fields bridge science, engineering and humanities, in contrast with the first wave, dominated by an engineering approach. The predominant engineering nature of the first experiments in automatic translation and the naive statistical linguistics in the fourties and fifties were to some extent a consequence of this fact.

The third wave, in the sixties, includes the emergence of the theory of programming languages (Ginsburg-Rice, Floyd), of the algorithmic theory of information (Kolmogorov-Chaitin), unfortunately still ignored by most linguists, of systems with incomplete information, of probabilistic grammars and automata, of the theory of fuzzy sets and, last but not least, of semiotics. Under the influence of this rapid development, the articles of mathematical linguistics are published in an increasing variety of journals: The Finite String (U.S.A.), Zeitschrift fur Phonetik, Sprachwissenschaft und Kommunikationsforschung (G.D.R.), Revue francaise du traitement de l'infor-

mation (France), Prague Bulletin of Mathematical Linguistics, while most review journals try with difficulty, and not always successfully, to adapt to the new situation; in both Mathematical Reviews and Zentralblatt fur Mathematik, in the section of computer science, the subsection of mathematical linguistics is called "Linguistics", despite the fact that most articles reviewed there are not concerned with natural languages; in the Russian journal Referativnyi Zurnal, mathematical linguistics is under the title "Matematicheskie problemy semiotiki".

In the same period, linguistics begins to be faced with many types of imprecision, other than randomness: fuzziness, genericity, typicality, approximation, and later roughness; even ambiguity appears in a new light; this remains a big challenge for quantitative linguistics too, where all these types are reduced implicitly to randomness, in absence of adequate tools. It is enough to recall that wellformedness still does not have a satisfactory status in respect to imprecision. All these problems attenuate considerably the border between quantity and quality. At the same time, "quantitative" interferes more and more with "computational", "cognitive", and mainly with "complexity"; the last one seems to be the most challenging paradigm in respect to "quantity". This hesitation in separating various approaches is well reflected in the diversity of journals dealing with the mathematical approach to language: "Theoretical Linguistics", "LInguistics and Philosophy", "Computational Linguistics", "Quantitative Linguistics" and the others already mentioned (besides them, many journals of computer science, of cognitive science, of artificial intelligence etc.).

The strong qualitative aspect of the problems today included in "Quantitative Linguistics" is also reflected in the "Call for papers" of the Third International Conference on Quantitative Linguistics and shows to what extent we have to take this label "cum grano salis".

Investigating Causal Relations in Language with the Help of Path Analysis

Juhan Tuldava, University of Tartu
Dr. Phil., Prof. Emer.
Aardla 9a-50, EE2481 Tartu, Estonia

Topical area: Methodological problems, model construction.

Summary. The paper presents the results of an experiment on the use of the method of Path analysis in identifying and measuring the causal relations in a system of linguistic objects. A short survey is given on the main principles and techniques of Path analysis.

Introduction

Path analysis, originally developed by S. Wright (1923; 1934), is a technique for evaluation of entire causal models. It uses a series of multiple regressions which, when combined, enable us to find out mutual relationships in a causal model by determining the magnitude of not only direct effects but also indirect effects. In this way, Path analysis makes it possible to assess the relative importance of different sources of causality, to partition the combined effects of the causal variables into mutually exclusive and meaningful components. Path analysis imposes a number of requirements on the relationships between the included variables, such as linear relationship between independent and dependent variables, dominant one-way ("simple recursive") relationship, absence of strict collinearity between the independent variables, uncorrelated residual factors (cf., e.g, Bohrnstedt & Knoke 1994).

In the past 20 - 30 years the model and methodology of Path analysis have been extensively studied and they have gained wide currency among social researchers. But the method of Path analysis has as yet found neither appreciation nor application in quantitative linguistics. The aim of this paper is to introduce the theory and techniques of Path analysis, illustrating them by the analysis of a linguistic causal system.

Initial data and structural equations

As an illustration for the application of Path analysis to linguistic material we shall analyze the multivariate distribution of some quantitative characteristics of words on the basis of a frequency dictionary of lexemes of non-conversational material taken from contemporary Estonian prose fiction. A fragment of the frequency dictionary, *1,200 most frequent words* (covering 75 % of the text), has been examined with regard to the following quantitative linguistic features, distributed to 12 frequency zones of 100 words each (see Tuldava 1995, Ch. 2, Table 1): (1) "age" of words (*A*), expressed by a coefficient denoting the ratio of ancient words (from the period before A.D. 1200) in the given frequency zone; (2) mean frequency (*F*) of words; (3) mean length (*L*) of words in syllables; (4) mean number of meanings, i.e. semantic scope (*S*), or polysemy of words in the given frequency zone.

In our hypothetical causal proposition we suppose that variables *A* (age) and *F* (frequency) cause *L* (word length) and all of them (*A*, *F*, *L*) cause *S* (polysemy). In terms of multiple linear regression, the relations among the variables in our model can be computed and represented by two regression equations:

$L' = 3.087 - 1.498\ A + 0.00008\ F$ and

$S' = -0.729 + 6.166\ A + 0.01\ F + 0.328\ L.$

In order to compare the relative magnitude of the regression coefficients we will have to *standardize* the coefficients for them to express how many standard deviations the dependent variable rises for one standard deviation increase in the independent variable.

This form of standardization is done by multiplying the regression coefficient expressed in the original measurement units by a fraction consisting of the standard deviation of the independent variable divided by the standard deviation of the dependent variable. For example, the first coefficient in the first equation above (-1.498) is standardized as follows (when $s_A = 0.1835$ and $s_L = 0.2896$; see Tuldava 1995: 19):

$(-1.498)(0.1835/0.2896) = -0.949.$

Because the mean of a standardized variable equals *zero*, the intercept in a standardized regression is also *zero*.

Computation gives us the following equations where the variables are measured not in terms of their original units but in standardized *z-scores*:

$z_L' = -0.949\ z_A + 0.031\ z_F$ and

$z_S' = 0.557\ z_A + 0.546\ z_F + 0.047\ z_L.$

The standardized coefficients ("Beta weights"), when used in Path analysis, are called *path coefficients*.

They are usually signified with the capital letter *P* with subscripts where the first subscript is always the dependent variable (I) followed by the independent variable (J): P_{ij}.

The Path model

Figure 1 displays the causal relations among the four variables, i.e. linguistic features - age (*A*), frequency (*F*), word length (*L*), and polysemy (*S*) under examination.

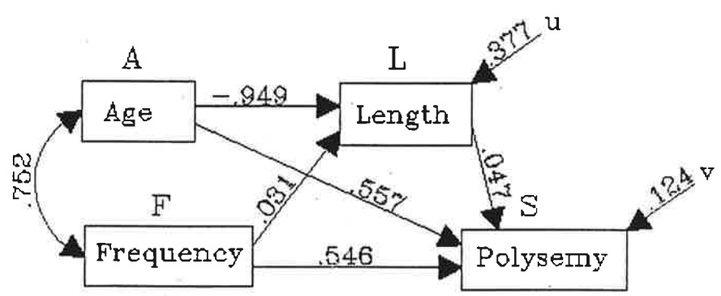


Fig. 1. Path diagram representing a model of causal relations in vocabulary

The variables placed on the utmost left of the diagram (*A* and *F*) are considered to be *exogenous*, or predetermined variables which remain unspecified, unanalyzed in our model. In the diagram they are linked by a curved double-headed arrow, indicating that they are related but not causally connected (In fact, the causal nature of the relationship is considered to be irrelevant for our purposes here; the correlation between *A* and *F*, i.e. r_{AF} equals 0.752).

The variables that are not exogenous are *endogenous* (here *L* and *S*) presumed to be controlled by other variables. The causes of their variations are represented within the model. The tail of a single-headed straight arrow emerges from the causal variable and the arrowhead points at the effect variable.

The path coefficients indicate the strength of the direct causal effect between the variables. Path coefficients also permit the calculation of the indirect causal effects through multiplication of the path values of the compound paths connecting two variables via intervening variables. The variables *u* and *v* in the diagram are called *residual variables*. They represent all the variables that have not been specified in the model but still have an effect on the dependent variables and reflect the amount of variation in these variables that has been unexplained.

Interpretation

The analysis and interpretation of the path model includes *decomposition*, or separation of a causal chain into its components.

For instance, the contribution from *A* to *S* is split into several components or paths which represent (see Fig. 1):

(1) a *direct* effect (symbolically, A→S): $P_{SA} = 0.557$;

(2) an *indirect* effect through *L* (A→L→S): $P_{LA}P_{SL} = (-0.949)(0.047) = -0.045$;

(3) *correlated* effects through *F* (A→F→S): $r_{AF}P_{SF} = (0.752)(0.546) = 0.411$

and through *F* and *L* (A→F→L→S): $r_{AF}P_{LF}P_{SL} = (0.752)(0.031)(0.047) = 0.001$.

When all actual figures are substituted into a structural equation, the components of the relationships between *A* and *S* will add up to the *total correlation coefficient*:

$r_{SA} = 0.557 - 0.045 + 0.411 + 0.001 = 0.924$.

Analogously, we can decomposite the other causal chains in our model as follows:

A→L: $-0.949 + (0.752)(0.031) = -0.166$;

F→L: $0.031 + (0.752)(-0.949) = -0.683$;

F→S: $0.546 + (0.031)(0.047) + (0.752)(-0.949)(0.047) = 0.709$.

The strongest total correlation seems to characterize the causal relation between *A* (age) and *S* (polysemy) with $r_{SA} = 0.924$, the direct effect being only 0.557. It was the

correlated effect (0.411) through the association with *F* (frequency) that played the decisive role in constituting the strong overall correlation. Thus old words, especially in combination with the frequency of occurrence, tend to be(come) polysemantic, which has been quantitatively evaluated in Path analysis.

While comparing, we can see that the substantial contribution from *F* (frequency) to *L* (word length) with $r_{LF} = -0.683$ is ascribable to the association with *A* (age). Obviously, shortening of words needs not only frequent use but also time. The same can be said about the contribution from *F* (frequency) to *S* (polysemy) with $r_{SF} = 0.709$ which consists of a direct effect of 0.546 and to a large extent of a correlated effect through the association with *A* (age): $(0.752)(0.557) = 0.195$.

Path analysis also allows us to spell out the relationships among the endogenous variables. Let us partition the correlation between *L* (word length) and *S* (polysemy) into its components. First, there is the direct effect of *L* on *S* (0.047). Second, *L* and *S* are associated to some extent because both are determined by *A* (age) and *F* (frequency). These *spurious* elements are represented by $(-0.949)(0.557) = -0.529$ and $(0.031)(0.546) = 0.017$.

There is also an association due to *related causes*. A cause of *L* (namely, *A*) is associated with a cause of *S* (namely, *F*), and this component of r_{SL} is represented by $(0.752)(-0.949)(0.546) = -0.390$. Also, another cause of *L* (namely, *F*) is associated with another cause of *S* (namely, *A*) and this is represented by $(0.752)(0.031)(0.557) = 0.013$. In sum $0.047 - 0.529 + 0.017 - 0.390 + 0.013 = -0.842$, which characterizes the total correlation between *L* and *S* (consisting of a direct effect, spurious elements, and related causes).

Conclusion

Thus, when all possible relations between the variables are included in the Path model, it becomes possible to see how much each of the various components (*direct effect*, *indirect effect*, *correlated effect*, *spuriousness*, *related causes*) contributes to the total correlation. Path analysis is, therefore, considered an important theoretical tool which forces the researcher to specify all the relationships in a causal model. Of course, we must remember that in organizing a causal model for Path analysis the researcher's

theoretical understanding - knowledge of linguistic relationships, logical deduction, etc. play the decisive role.

Our simple example does not touch upon all the problems connected with Path analysis. If we were to pursue the research, we would want to specify more elaborate models, including additional possible causes and using other versions of Path analysis which would enable us to cope with such problems as nonrecursiveness, multicollinearity, latent variables and others (see, e.g., Blalock 1971; Wold 1974; Heise 1975). As it was our first attempt at introducing Path analysis to quantitative linguistics, we restricted our investigation to the most simple model of Path analysis. (Looking for criticism of the theory and method of Path analysis, see, e.g., McPherson 1990: 594-595).

Note. All computation in this experiment has been carried out by the CALIS procedure (SAS package).

References

Blalock, H. M. Jr. (ed.) (1971). *Causal models in the social sciences.* Chicago: Aldine-Atherton.

Bohrnstedt, G.W. & Knoke, D. (1994). *Statistics for social data analysis.* 3rd ed. Itasca, Ill.: Peacock.

Heise, D.R. (1975). *Causal analysis.* New York: Wiley.

McPherson, G. (1990). *Statistics in scientific investigation.* New York: Springer.

Tuldava, J. (1995). *Methods in quantitative linguistics.* Trier: Wissenschaftlicher Verlag Trier.

Wold, H. (1974). Causal flows with latent variables. *European Economic Review*, **5**, 67-86.

Wright, S. The theory of path coefficients. *Genetica*, **8**, 239-255.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, **5**, 161-215.

V.V.NALIMOV'S PROBABILISTIC MODEL OF LANGUAGE AND CONSCIOUSNESS

Vladimir A.Dolinsky, Ph.D.; Zhanna A.Drogalina

V.Dolinsky, Moscow State Linguistic University. Department of applied and experimental linguistics. Docent. 129345, Moscow, Ostashkovskaya, 9-2-98. Russia. Tel.: /095/ 475 8384. E-mail: nalimov@Nalimov.home.bio.msu.ru

Zh.Drogalina, Moscow State University. Biological Department. Researcher. 117415, Moscow, Udaltsova, 4-327. Tel.: /095/ 131 4530.

Philosophy of language and conscoiusness elaborated by outstanding mathematician Vasily Vasilyevich Nalimov, author colleague, tutor and guru, is regarded.

TOPICAL PAPER. Epistemological issue on explanation of language phenomena; study of consciousness; philosophy of science; model construction; quantitative methods.

On January 19, 1997, Dr.Nalimov, 86 years old, died in his Moscow flat. Born in Moscow in 1910 and educated as a mathematician, he has had spiritual teachers from the esoteric tradition of Mystical Anarchism in Russia and spent eighteen years (1936-1954) in GULAG. He has also worked (nearly ten years) with the academician Andrey N.Kolmogorov, who is known world-wide as a genius in mathematics. A mathematician and philosopher, Dr. of Technology and Professor of Mathematical Statistics and theory of Probability, Nalimov was also elected academician in september 1996. His theory of "probabilistic way of mind" has been presented as a series of books published in Russia and abroad.

Angela Thompson (1993) in article, which have been pub-

lished in leading American psychological journal, calls Nalimov a "Russian visionary" and quotes Stanislav Grof and David Bohm. Thompson believes that "Nalimov's concept of meaning and consciousness, which will probably not be fully realized untill well into the next century, encompass topics as varied as language, mathematics, and philosophy" (pp.82-83).

According to V.V.Nalimov, the great interest of philosophers in language problems can be easily explained: the study of language is a way of studying thinking. It seems reasonable to believe that epistemology may be turned from a theoretical-speculative subject into a natural science if language is made an object of study. Then it will become possible in epistemology to discuss hypotheses in comparison with actually observed phenomena, as is the case in other natural sciences. When an experiment is performed, it will be possible to express the results of some observations regarding language in quantitative terms, and hypotheses will be verifiable. In such an approach the profoundness in problem formulation which is characteristic of classical epistemology is lost.

The same, however, happened in physics: classical physics was to a significant degree purely metrological (i.e., based on measurements and their interpretetion) and remained far from philosophical analysis. Modern physics, with such sections as quantum mechanics and the theory of relativity, looks quite different. Here, philosophical problems are already touched upon, but, again, there are no longer so pro-

found as they used to be in traditional philosophy. Reading even very serious basic papers of modern physics, we still do not learn anything about the general philosophical state of mind of their authors. The same happens when we read papers of philosophers about language.

Nalimov's main aspiration and striving is an attempt to elaborate a philosophical background of a probabilistically oriented world outlook. Most important outcome of the work he thinks to be:

1. Creation of a school of mathematical methods of experimental design.

2. Formulation of the conception of Scientometrics, including the coining of the very term.

3. Elaboration of probabilistically oriented model of language, consciousness and evolution viewed as a self-organization process.

4. Critical analysis of the situation in modern science. Raising the problem of "scientific" in science to show that modern science fails to meet the requirements to be "scientific" as it was formulated in the past.

5. Elaboration of the integrated world outlook based on Plato's philosophy as an attempt to return to the philosophical classics. Formulating the premises for the mathematical model of consciousness. Constructing the probabilistic logic, exampted from the law of the excluded middle. The model is justified by its heuristic power. Here Nalimov appears to be close to the school of intuitionistic mathematics (particularly L.E.J.Brouwer) which favors intuitive cons-

tructions.

Principal positions by V.V.Nalimov are:

1. Philosophical and linguistic conception should be based on axioms (premises).

2. Philosophy has to be developed in close relationship with sciences. Most ideas with philosophical background come from such fields of science as mathematics, physics, cosmogony, biology and non-traditional psychology (such as transpersonal and humanistic psychology).

3. At the same time there should be preserved the connection with classical thought rooted in ancient Greek and Eastern conceptions of man and Universe.

4. It is natural to apply to mathematical constructions, as human consciousness is provided with the ability to contact the world through mathematical forms and categories such as spase (with the variety of geometries), time (according to the present conceptions), number (which nature is non-material), probability, and attached to it - spontaneity and freedom (!).

5. There is a Mystery in the Universe - unfolding our knowledge; we do not destroy it but expand and deepen its image. After all, science is exposing but an elaborated Un-knowledge, which looks now much more serious than it was in past times.

An axiomatic system of a probabilistically oriented theory of language and consciousness is based on the formula by Bayes, previously used only in mathematical statistic. Premises to the point are the following:

1. All potential meanings (Russian word "smysly") of the world are primordially given (same way as fundamemtal constants are given in the physical world).

2. All meanings are initially correlated with the linear continuum of Cantor (otherwise, the meanings of the world are compressed the way numbers on the real axis 'm' are).

3. Compressed meanings represent the unpacked (unmanifested) World - so called semantic vacuum.

4. Unpacking (emergence of texts) is realized through probabilistic weighning of the axis 'm' - different measures are ascribed to its different intervals (the metric of the scale 'm' is assumed to be initially given and remaining unaltered).

5. Any change in the text - its evolution - is linked with a spontaneous emergence in a situation 'y' of the filter p(y/m) that interacts multiplicatively with the initial function p(m). The interaction is given by the well-known formula by Bayes:

$$p(m/y) = kp(m)p(y/m)$$

where the distribution function p(m/y) determines the semantics of a new text emerging after the evolutionary impetus 'y'; 'k' is a normalizing constant. In our case the Bayesian formula acts as a syllogism: from the two premises p(m) and p(y/m) necessarily follows a text with new semantics p(m/y). In Bayesian syllogism, in contrast to the categorical syllogism by Aristotle, both the premises and the corollary are not atomic but probabilistically fuzzy, and at least the se-

cond premise is of a conditional (conditioned by the situation 'y') character, but not a categorical one.

New Bayesian-Nalimovian logic, applied to working out the problem of language, consciousness and culture, allows:

- to comprehend and evaluate word polymorphism of ordinary (everyday) language; it is due to semantic polymorphism we are free from the consequences of Godel's theorem while narrating something (before became an abstract concept, a sign submited to logic of discourse, the word should be interpreted on a level of uncounscious lingual symbolism);
- to explain varieties of text comprehension;
- to describe the emergence of texts;
- to analyze the problem of spontaneity in word association process;
- to grasp such hard notions as freedom, nirvana, spontaneity, etc.;
- to view the human Ego as a specific - self-interpreting text;
- to analyze the problem of transcendency - going beyond the limits of personal consciousness;
- to make a semantic account of biological evolutionism;
- to accept the idea of ubiquity of at least weak forms of consciousness.

Nalimov states that such semantic manifestation as space, time, number, probability, spontaneity, are inherent in the Universe and independent of the presence of the observer.

The works by Nalimov are quite complicated due to their interdisciplinary character. He was sure that most troubles of modern situation are rooted in the alienation from the Culture, which is not experienced as an integrated wholeness, but fragmentarily. V.V.Nalimov assumed that philosophy nowadays should challenge over-technologization of Culture and release it from decrepit mechanistical conceptions.

Literature

Dolinsky, V.A. 1995. Linguistic modelling and extralinguistic meanings // Lingvistika na iskhode XX veka. T.1. Moscow, pp.159-161.

Dolinsky, V.A. 1995. Oblomok korablekrusheniya <L'epave> (A dialogue with V.V.Nalimov) // Sila Dukha, N 4, pp.4-11.

Nalimov, V.V. 1969. Naukometriya <Scientometrics>. Moscow: Nauka, 192 p. The book was also issued in Poland, 1971, and in Hungary, 1980.

Nalimov, V.V. 1975. Theorie des Experiments. Berlin: VEB Deutscher Landwirtschaftsverlag. 159 S. In Russian: Nalimov, V.V. 1971. Teoriya experimenta. Moscow: Nauka, 207 p.

Nalimov, V.V. 1981. In the labyrinths of language: A mathematician's journey. Philadelphia, PA: ISI Press, 246 p. In Russian: Nalimov V.V. 1979. Veroyatnostnaya model' yazyka, 2nd edition, 303 p. (1st.edition 1974). The book was also published in Poland in 1976.

Nalimov, V.V. 1982. Realms of the unconscious: The enc-

hanted frontier. Philadelphia, PA: ISI Press, 320 p. In Russian: Nalimov, V.V., Drogalina, Zh.A. 1995. Real'nost' nereal'nogo. Moscow: Mir Idey, 420 p. In French: Nalimov, V.V. 1996. Les Mathematiques de l'Inconscient. Monaco: Editions de Rocher. 488 p.

Nalimov, V.V. 1989. Can Philosophy be Mathematized? Probabilistic Theory of Meaning and Semantic Architectonics of Personality // PHILOSOPHIA MATEMATICA, An International Journal for the Philosophy of Modern Mathematics, II, v.4. (Also published in Poland, 1990, and in Russia, 1991).

Nalimov, V.V. 1989. Spontannost' soznaniya <Spontaneity of consciousness>. Moscow: Prometei. 287 p.

Nalimov, V.V. 1992. Spontaneity of Consciousness. An Attempt of Mathematical Interpretation of Certain Plato's Ideas // Carvallo M.E. (Ed.) NATURE, COGNITION AND SISTEM II. Dordrecht: Cluwer Press.

Nalimov, V.V. 1993. V poiskakh inykh smyslov. Moscow: Progress. 261 p.

Nalimov, V.V. 1994. Na grani tretyego tysyacheletiya. Moscow: Labirint. 73 p.

Nalimov, V.V. 1994. Kanatokhodyec <A Rope-Dancer> (A memoir text). Moscow: Progress. 454 p.

Nalimov, V.V. 1996. Kritika istoricheskoy epokhi: neizbezhnost' smeny kultury v XXI veke // Voprosy filosofii, N 11, pp.66-76.

Thompson, A.M. 1993. Vasily Vasilyevich Nalimov: Russian visionary // Journal of Humanistic Psychology, v.33, N 3, pp.82-98.

The parameters of LNRE models as sample-size invariant: problems and opportunities

R. Harald Baayen, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands; e-mail: baayen@mpi.nl.
Fiona J. Tweedie, University of Glasgow, University Gardens, Glasgow, G128QQ, United Kingdom; e-mail: fiona@stats.gla.ac.uk.

Topical paper
Topical area: lexical statistics, quantitative stylistics

Abstract

Many characteristic text constants are heavily dependent on the sample size. To avoid this dependence, (LNRE) models have been developed, the parameters of which are, in theory, sample-size invariant. In practice, the parameters of LNRE models may nevertheless reveal considerable dependence on the sample size. We show that this is a direct consequence of a lack of goodness-of-fit, and can therefore be used for goodness-of-fit testing. This dependence, furthermore, requires new techniques for cross-text comparisons. We propose one such method, based on the developmental profiles of LNRE parameters.

1 Introduction

A well-known problem in the domain of lexical statistics concerns the dependence of a many measures of lexical richness on the sample size N, the number of word tokens included in a corpus or text-based frequency count (see Tweedie and Baayen, 1997, for a review). It is well-known that the free parameter of Zipf's law in its original forms (Zipf 1935, 1949) is similarly subject to this dependence, as shown by Orlov (1983a, 1983b). Orlov and Chitashvili (1982a, 1982b, 1983a, 1983b) developed Zipf's law into a fully-fledged Large Number of Rare Events (LNRE) model in which the dependence of lexical distributions on N is accounted for in a principled way by means of an additional parameter Z, the unique sample size for which Zipf's law in its simple form holds. Other models, such as the lognormal model (Carroll 1967) and the inverse Gauss-Poisson model (Sichel 1986) likewise belong to the class of LNRE models (Chitashvili and Baayen, 1993).

The parameters of LNRE models are **in theory** invariant with respect to the sample size. The problem addressed in this paper is that, to our dismay, **in practice** the parameters of LNRE models may nevertheless reveal substantial dependence on N. In section 2 we focus on the sources of this systematic variation in the values of LNRE parameters. The empirical dependence of LNRE parameters on the sample size requires special methods when texts of different lengths are to be compared. In section 3 we propose such a method, based on the comparison of the empirical developmental profiles of LNRE parameters.

2 LNRE parameters and sample size

The upper panels of Figure 1 show that the parameter Z of the extended Zipf's law (Orlov 1983a) and the parameter b of the inverse Gauss-Poisson law (Sichel 1986) are no exception to the observation that most measures advanced as independent of the text length N tend to vary systematically with N. The horizontal axes of Figure 1, which is based on L. Carroll's *Alice's Adventures in Wonderland*, display the sample size N. The vertical axes of the upper panels display the values of Z (left) and b (right) as a function of N. The dots show the observed, empirical values of these parameters as estimated for the sequence of 20 equally-spaced text lengths $N = 1326, 2652, \ldots, 25180, 26505$ on the basis of the frequency spectra at these points in 'sample time'. For the extended Zipf's law, we find that the estimated value of Z increases with N. For the inverse Gauss-Poisson model, b also reveals considerable variation, especially for small N. The estimates of b, however, tend to converge relatively quickly to its final value as estimated for the complete text.

The solid lines in the upper panels of Figure 1 show the expected values of Z and b as calculated on the basis of a series of Monte Carlo randomizations of the words in *Alice's Adventures in Wonderland*. We observe a clear pattern of dependence on the sample size N. In the case of Z, we see an initial steep decline,

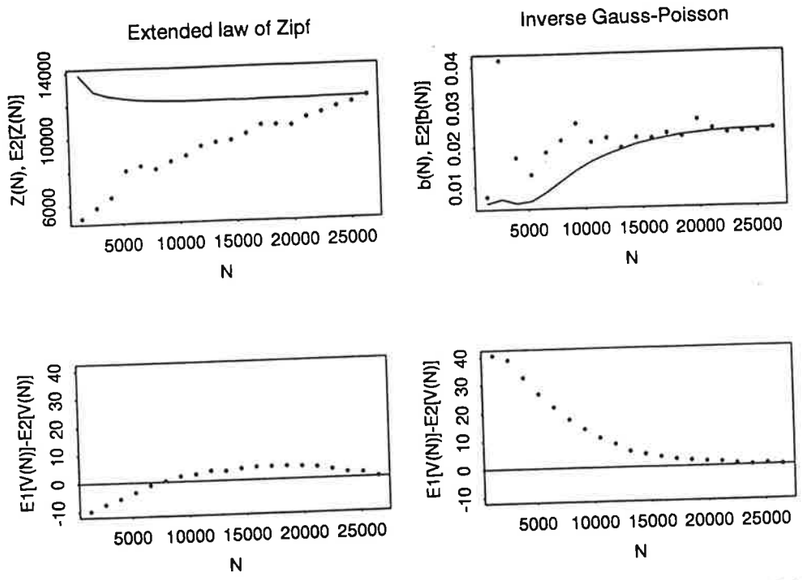


Figure 1: *Observed (dots) and Monte-Carlo-based expectations (solid lines) for Z (upper left) and b (upper right), and the error for the vocabulary size for the extended Zipf's law (bottom left) and the inverse Gauss-Poisson law (bottom right) as calculated for* Alice's Adventures in Wonderland, *measured at 20 equally-spaced intervals. E1: expectation based on the model for the full text ($E_{LNRE}[V(N)]$); E2: Monte Carlo expectation ($E_{MC}[V(N)]$).*

after which Z stabilizes, albeit with some very small concave curvature. In the case of b, we find a convex function that levels off by the end of the text.

Why do the Monte Carlo expectations reveal a non-random developmental profile as a function of N. Does this imply that LNRE models fail to eliminate the ubiquitous dependence on N which they were designed to overcome, not only in practice, but also in theory?

Fortunately, this conclusion is unwarranted. The changing values of Z and b are a direct consequence of imperfections in the fit of the LNRE models to the empirical frequency spectrum of *Alice's Adventures in Wonderland*. To see this, consider the bottom panels of Figure 1, which plot the difference between two theoretical growth curves, $\mathrm{E}_{LNRE}[V(N)]$ and $\mathrm{E}_{MC}[V(N)]$. The expected vocabulary growth curve $\mathrm{E}_{LNRE}[V(N)]$ is obtained by estimating the parameters of the models for the complete text ($N = 26505$), followed by interpolation of the expected vocabulary size for 20 equally-spaced smaller sample sizes using

$$\mathrm{E}_{LNRE}[V(N)] = \frac{Z}{\log(p^* Z)} \frac{N}{N-Z} \log(N/Z) \tag{1}$$

for the extended Zipf's law (with p^* the maximum relative frequency in the text), and using

$$\mathrm{E}_{MC}[V(N)] = \frac{2}{bc}\left[1 - e^{b(1-\sqrt{1+Nc})}\right] \tag{2}$$

for the inverse Gauss-Poisson law (with c the second parameter of the model, and fixing its third parameter, γ, at -0.5 for computational tractability). The second expected vocabulary growth curve, $\mathrm{E}_{MC}[V(N)]$, is obtained by calculating the average vocabulary size in a series of randomizations, i.e. Monte Carlo expectations, for the same 20 measurement points in sample time. Since both expectations are based on the urn model, their values should be identical. Hence, their difference $\mathrm{E}_{LNRE}[V(N)] - \mathrm{E}_{MC}[V(N)]$ is a diagnostic for how well an LNRE model fits the data.

For the extended Zipf's law, we observe a convex curvature. For small N, the model underestimates $V(N)$, for medium N, it reveals a slight overestimation bias compared to the Monte Carlo expectations. Since increasing Z leads to an increase in $\mathrm{E}_{LNRE}[V(N)]$ by (1), the underestimation bias of $\mathrm{E}_{LNRE}[V(N)]$ observed for small N is compensated for by increasing Z when estimating this parameter for small sample sizes in the randomizations. For larger sample sizes, the mismatch between $\mathrm{E}_{LNRE}[V(N)]$ and $\mathrm{E}_{MC}[V(N)]$ is so small that the value of Z is hardly affected, and approaches constancy.

Turning to the inverse Gauss-Poisson law, we find an overestimation bias for $\mathrm{E}_{LNRE}[V(N)]$ that decreases with increasing N. This model accommodates its overestimation bias by increasing c and by decreasing b as N becomes smaller. Compared to Z, the value of b becomes reasonably stable at a rather late moment in sampling time N. This is due to the larger bias of $\mathrm{E}_{LNRE}[V(N)]$ for the inverse Gauss-Poisson model. Consequently, greater changes in the parameters are required to accommodate the model to the structure of the frequency spectra of the smaller sample sizes in the Monte Carlo simulations.

Since the observed dependence of LNRE parameters on the sample size directly reflects the accuracy of LNRE models, we can make use of this dependence to evaluate the goodness-of-fit of these models. Traditionally, the goodness-of-fit of LNRE models is evaluated by means of chi-square tests. Unfortunately, the appropriate chi-square test (using the covariance matrix of the spectrum elements) almost always leads to the rejection of theoretical models with p-values that may be as small as 10^{-8} (see also Grotjahn and Altmann, 1993), even when fits are obtained that are perfectly reasonable to the eye. Instead of using the chi-square test, the extent to which LNRE parameters change as a function of N can be used as a measure of goodness-of-fit: the less accommodation required, the better the fit of the model. As a practical measure, we propose to use the percentage of measurement points for which the absolute error $|\mathrm{E}_{LNRE}[V(N)] - \mathrm{E}_{MC}[V(N)]|$ falls below a given tolerance threshold δ:

$$D(K,\delta) = \frac{1}{K}\sum_{i=1}^{K} \mathrm{I}\left[|\mathrm{E}_{LNRE}[V(N_k)] - \mathrm{E}_{MC}[V(N_k)]| < \delta\right], \tag{3}$$

with K the number of measurement points (20 in our examples) and $V(N_k)$ the vocabulary size at the k^{th} measurement point. We choose the model error δ as small as possible, but such that the proportions $D(K,\delta)$ for the two models differ significantly. For the extended Zipf's law and the inverse Gauss-Poisson law, the smallest significant difference is found for $\delta = 5$: $D(20,5) = 0.85$ for Z, and $D(20,5) = 0.5$ for b ($p < 0.05$). These calculations formalize the visual impression of Figure 1 that the extended Zipf's law provides the better fit to *Alice's Adventures in Wonderland*, even though it has only one parameter to vary instead of two.

3 Comparing developmental profiles of LNRE parameters

The above test for goodness-of-fit pits the predictions of LNRE models against the predictions of the urn model (without replacement). However, words do not occur randomly in texts. As illustrated in the upper panels of Figure 1, the observed values of the parameters diverge from their Monte Carlo expectations for a wide range of measurement points. This is due to the non-random, underdispersed use of words in discourse, which, in the case of *Alice's Adventures in Wonderland*, causes the empirical vocabulary size to be substantially smaller than its theoretical expectation for all 20 measurement points (see Baayen, 1996, for detailed discussion). In the case of the extended Zipf's law, the overestimation bias of the theoretical estimates is compensated for when we estimate Z for smaller text lengths using (1). In order to match the expected and the observed vocabulary size, we have to lower $\mathrm{E}_{LNRE}[V(N)]$ compared to what we would expect given the complete text, and hence Z has to be lowered too. In the case of the inverse Gauss-Poisson law, the parameters b and c are likewise adjusted to meet the requirement that for each measurement point the expected vocabulary size should be equal to its expectation given the frequency spectrum at that measurement point.

For practical applications in quantitative stylistics, in which texts of different lengths may have to be compared, the observed functional dependence of empirical LNRE parameters on the sample size N reintroduces the problem of how this dependence should be taken into account. Figure 2 shows how severe this problem is. For a random selection of electronically available texts, it plots the empirical values of Z and b, when estimated from successively larger samples, and shows that they tend to change systematically

with N. Since the values of Z estimated for the full text sizes are not characteristic for the text as a whole, it seems reasonable to base between-text comparisons on the developmental profiles $Z(N)$ and $b(N)$. To do so, we may proceed as follows.

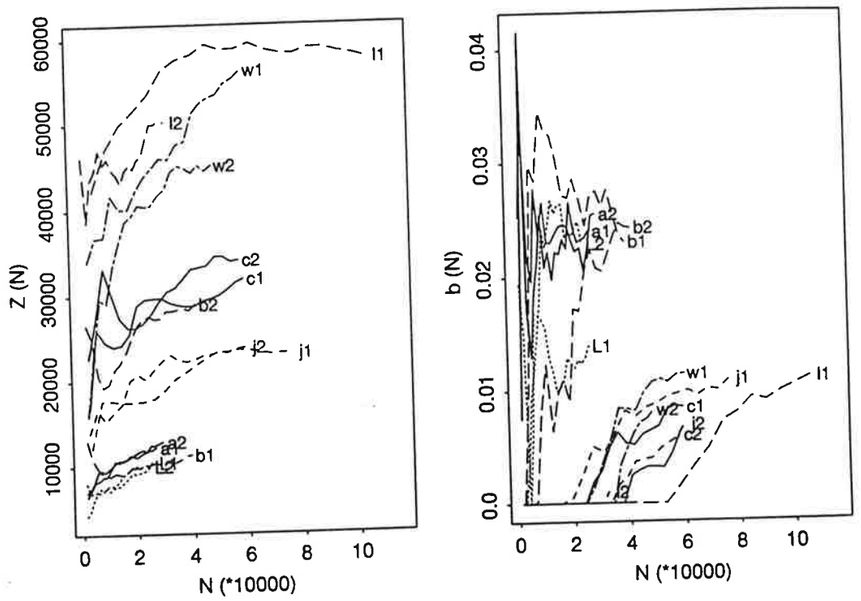


Figure 2: *Empirical values of Z and b as a function of N in selected texts. a1, a2: Carroll; b1, b2: Baum; c1, c2: Conan-Doyle; j1, j2: James; l1, l2: London; L1, L2: Luke-Acts (KJV); w1, w2: Wells. See the Appendix for bibliographical details.*

We will fit a linear model to the data, treating the K measurements from each text as repeated measures of the same value. Our model will be

$$X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \tag{4}$$

where X_{ij} is the value of either Z or b in text i at measurement point j, here $i = 1, \ldots, 14$ and $j = 1, \ldots, 20$. In order that the analysis can be carried out, α_1 and β_1 are constrained to equal 0. The ϵ_{ij} terms are error terms, with the usual normality, independence and zero mean assumptions. Note that we are comparing the texts at each measurement point, i.e. the first recorded value of $Z, Z_{1,1}$ is compared with other first values of $Z, Z_{2,1}, \ldots, Z_{14,1}$, regardless of the values of N_1 in each text.

We fitted the model in (4) to the observed values of Z_{ij} and b_{ij} and examined the residual plots; the model seemed to be a good fit for the data, with a small number of rather high residuals. An Analysis of Variance (ANOVA) table reveals that for Z, there are significant differences between repeated measures factors ($p = 0.0020$ using the Greenhouse-Geisser Most Conservative Test adjustment to the degrees of freedom in the F-test) and text factors ($p < 0.0000$). The multiple R^2 value is 97.47%, indicating that almost all the variation in the observed data is explained by the factors. A similar analysis for b produces similar results for between text ($p < 0.000$) and repeated measures ($p = 0.0260$ with the adjusted F-test) factors. Here $R^2 = 83.81\%$, indicating a slightly less good fit to the data, perhaps due to the unusual structure of the data, as shown in the second panel of Figure 2.

Given that there are reliable differences between the developmental profiles of the LNRE parameters, we may now proceed to investigate which of the factors α_i and β_j are significantly different from which others. We will present here the results of comparing values of α_i for both Z_{ij} and b_{ij}.

As there are almost 200 possible comparisons of the form $(\alpha_i - \alpha_{i'})$ we will use Bonferroni adjusted t-intervals. Even with this strong restriction on our confidence intervals the only factors that are not significantly different are those associated with the Conan Doyle texts (c1, c2), the James texts (j1, j2) and the St Luke texts (L1, L2). In addition, there is no significant difference observed between the text of *Through the Looking-Glass and what Alice found there*, text a2, and either of the Luke texts. Turning to the factors fitted to the model of b_{ij} we find that the only non-significant difference between α_is is to be found between *Alice's Adventures in Wonderland* and *The Acts of the Apostles*, the second of the texts by St Luke (L2).

It is clear from these comparisons that there is some evidence of within-author homogeneity in terms of Z_{ij}; texts by Conan Doyle, James and St Luke are not found to be significantly different. The lack of significance between texts by Carroll and St Luke can be explained by examining the first panel of Figure 2; the developmental profiles of Z are very similar for the Carroll and St Luke texts. Note, however, that the texts by Baum (b1, b2) illustrate that the developmental profiles of texts written by the same author can be far apart. While authorial structure is partially evident in the results for Z, it is entirely absent from the model for b. There is no indication that any of the texts might be by the same author, except for a Carroll and St. Luke text. It is clear that Z is better at elucidating authorship, although the model (4) clearly requires more investigation.

We have only presented the differences between factors for the author effect of this straightforward model. Examinations of the residuals in more detail may prove fruitful and analysis of the repeated measures factors would reveal change points in the vocabulary usage in the texts. It is also important to consider other models; a text factor could be nested under the author factor, or each text could be considered as a simple replication of the author's Z values. Other types of analysis could also be employed for this data; the temporal structure of the text could be taken into account in a time series analysis. Tweedie and Baayen (1997) consider confidence intervals for such parameters using partial randomisations of the texts.

4 Conclusions

We have called attention to the dependence of the parameters of LNRE models on the sample size. This dependence is observed for both the empirical development of a text, as well as for its theoretical development in Monte Carlo simulations. We have focused on the new possibilities that this finding offers for goodness-of-fit testing and for between-texts comparisons. An important problem that we have not touched upon is how LNRE models might be enhanced to take the non-random use of words in discourse and its effects on the accuracy of their predictions into account. We are currently investigating the possibility of using link functions that specify how a parameter such as Z changes over sampling time as a function of discourse structure. In this way, we hope to formulate LNRE models with enhanced empirical adequacy.

References

Baayen, R. H.: 1996, The effect of lexical specialization on the growth curve of the vocabulary. To appear in *Computational Linguistics* **22**.

Carroll, J. B.: 1967, On sampling from a lognormal model of word frequency distribution, *in* H. Kučera and W. N. Francis (eds), *Computational Analysis of Present-Day American English*, Brown University Press, Providence, pp. 406–424.

Chitashvili, R. J. and Baayen, R. H.: 1993, Word frequency distributions, *in* G. Altmann and L. Hřebíček (eds), *Quantitative Text Analysis*, Wissenschaftlicher Verlag Trier, Trier, pp. 54–135.

Grotjahn, R. and Altmann, G.: 1993, Modeling the distribution of word length: some methodological problems, *in* R.Koehler and B.B.Rieger (eds), *Contributions to quantitative linguistics*, Kluwer, Dordrecht, pp. 141–153.

Orlov, J. K.: 1983a, Dynamik der häufigkeitsstrukturen, *in* H. Guiter and M. V. Arapov (eds), *Studies on Zipf's Law*, Brockmeyer, Bochum, pp. 116–153.

Orlov, J. K.: 1983b, Ein model der häufigskeitstruktur des vokabulars, *in* H. Guiter and M. V. Arapov (eds), *Studies on Zipf's Law*, Brockmeyer, Bochum, pp. 154–233.

Orlov, J. K. and Chitashvili, R. Y.: 1982a, On some problems of statistical estimation in relatively small samples, *Bulletin of the Academy of Sciences, Georgia* **108**, 513–516.

Orlov, J. K. and Chitashvili, R. Y.: 1982b, On the distribution of frequency spectrum in small samples from populations with a large number of events, *Bulletin of the Academy of Sciences, Georgia* **108**, 297–300.

Orlov, J. K. and Chitashvili, R. Y.: 1983a, Generalized Z-distribution generating the well-known "rank-distributions", *Bulletin of the Academy of Sciences, Georgia* **110**, 269–272.

Orlov, J. K. and Chitashvili, R. Y.: 1983b, On the statistical interpretation of Zipf's law, *Bulletin of the Academy of Sciences, Georgia* **109**, 505–508.

Sichel, H. S.: 1986, Word frequency distibutions and type-token characteristics, *Mathematical Scientist* **11**, 45–72.

Tweedie, F. J. and Baayen, R. H.: 1997, *How Variable May a Constant be? Measures of Lexical Richness in Perspective*. Manuscript.

Zipf, G. K.: 1935, *The Psycho-Biology of Language*, Houghton Mifflin, Boston.

Zipf, G. K.: 1949, *Human Behavior and the Principle of the Least Effort. An Introduction to Human Ecology*, Hafner, New York.

Appendix

Author	Title	Key
Baum, L. F.	*The Wonderful Wizard of Oz*	b1
	Tip Manufactures a Pumpkinhead	b2
Carroll, L.	*Alice's Adventures in Wonderland*	a1
	Through the Looking-glass and what Alice found there	a2
Conan Doyle, A.	*The Hound of the Baskervilles*	c1
	The Valley of Fear	c2
James,	*Confidence*	j1
	The Europeans	j2
St Luke	*Gospel according to St Luke (KJV)*	L1
	Acts of the Apostles (KJV)	L2
London, J.	*The Sea Wolf*	l1
	The Call of the Wild	l2
Wells, H. G.	*The War of the Worlds*	w1
	The Invisible Man	w2

PHONETICS

PHONOLOGY

Coevolution of Phonology and the Lexicon in Twelve Languages of West Africa

Daniel Nettle
Merton College
Oxford
Great Britain
GB-OX1 4JD

Daniel.Nettle@merton.ox.ac.uk

Summary

Synergetic models of language structure predict that the length of a word will depend upon various parameters such as its frequency and the number of phonemes in the language. This prediction has been used to explain word length differences within languages, but less often to explain the differences *between* languages. Here I show that average word length across 12 West African languages is related to the phonological inventory. I outline a mechanism by which this relationship evolves.

Topical paper: Language typology, systems theory

Word Length: A Systems Theoretical Model

It has long been obvious that words have very different typical lengths in different languages, from the monosyllables of Chinese and other Asian languages, to the many-syllable roots of Hawaiian and the languages of Australia. Nonetheless, the word in its uninflected stem form is considered a basic, universal linguistic unit which is comparable across languages. The question thus arises of why words should be of such different composition in different languages.

Systems-theoretical linguistics treats a language as a dynamical, self-organising system whose structure is optimised to its function of communicating and representing information. The pressure exerted by function on structure is not uni-directional, however. Rather, languages evolve under several *competing motivations*, such as the minimisation of memory load and the minimisation of ambiguity, the minimisation of articulatory effort and the maximisation of acoustic distinctiveness. The interactions between these different pressures can be formalised into systems-theoretical models, and predictive statements about language structure produced (Köhler 1986, 1987).

According to Köhler's models, the length of a individual word will be a function of the number of segments in the phonological inventory of the language, the word's frequency, the number of words in the lexicon, and the degree of redundancy which the language requires due to the noisy nature of the human speech channel. Where L is the length of a word, then:

(1) $L = a\,(Segments^{b})(Frequency^{c})(Lexicon^{d})(Redundancy^{e})$

- where a,b,c,d and e are constants.

This statement about the distribution of word length within a language can be modified to make predictions about typical word length *across* languages. If we take a sample of words of all different frequencies from a language, and average their length, then their mean λ will be distributed as:

(2) $\lambda = a\,(Segments^{b})(Lexicon^{d})(Redundancy^{e})$

We can reasonably assume that the redundancy parameter is constant across words and across speech communities , given that the human speech mechanisms are everywhere the same. Furthermore, although there are differences in lexicon size between different languages, their effects will be negligible as long as all lexicons are large and d is small.

We can therefore take (Lexicon[d]) and (Redundancy[e]) as constants when comparing across languages, giving:

(3) $\lambda = a\ (\text{Segments}^{b})$

The prediction that there will be a relationship of the form given in (3) has already been tested and found to be correct for ten unrelated languages (Nettle 1995). In this paper, I repeat the analysis for twelve West African languages, and then investigate the mechanisms which lead to the synergetic relationship.

Testing the Prediction

Methods

The data used to test prediction (3) were gathered as part of a wider investigation of the areal linguistics of West Africa (Nettle 1996). The twelve languages were chosen for the quality of information available, and though they are all genetically or areally related in some way, the relationships are sufficiently uniform not to compromise statistical independence. To test the prediction, two data are needed for each language:

(i) *A phonological inventory*. This was obtained from published sources in all cases. The number of contrastive segments, henceforth S, was added up, using uniform criteria outlined in Nettle (1996). As the number of possible contrasts at each segmental slot was desired, each vowel/tone combination for tone languages was considered separately. Thus, a language with five vowels and three contrastive tones is deemed to have 15 possible contrasts on a vowel position (see Nettle 1995 for details of this method). The measure S is the simple sum of vowel and consonant contrasts, and therefore takes no account of phonotactic rules operating in the language. It is thus a simplification, whose accuracy depends upon the similarity of phonotactics from language to language.

(ii) *An estimate of the average word length* (λ). This was obtained by measuring the length in segments of randomly-sampled stems in a dictionary of the language. The results of Nettle (1995) suggested that a sample of 50 stems was sufficient, and that dictionary size was unimportant as long as all dictionaries contained more than 1000 entries.

Results

The number of segments in the inventory of the language (S), and the average word length (λ) are shown for each language in table 1.

Language	S	λ
Fula	33	6.42
Hausa	35	5.68
Tamasheq	36	5.26
Songhai	42	4.96
Bambara	49	4.86
Ngizim	52	5.32
Edo	53	4.42
Igbo	58	4.62
Mende	71	4.7
Ewe	81	4.16
Vata	164	4.56
Vute	195	3.94

Table 1 Inventory size (S) and average word length (λ) for the twelve languages. Sources as given in Nettle (1996)

There is indeed a negative power relationship between λ and S, as figure 1 shows. Curve estimation using the SPSS computer package gives the equation:

(4) $\lambda = 10.18\ S^{-0.18}$

($r^2 = 0.60$, $d.f. = 10$, $p < 0.001$)

The prediction of the systems-theoretical model is thus met, and the finding of Nettle (1995) for ten unrelated languages replicated. If the two data sets are combined, the following overall equation is produced:

(5) $\lambda = 17.87\ S^{-0.31}$

($r^2 = 0.67$, $d.f. = 20$, $p < 0.001$)

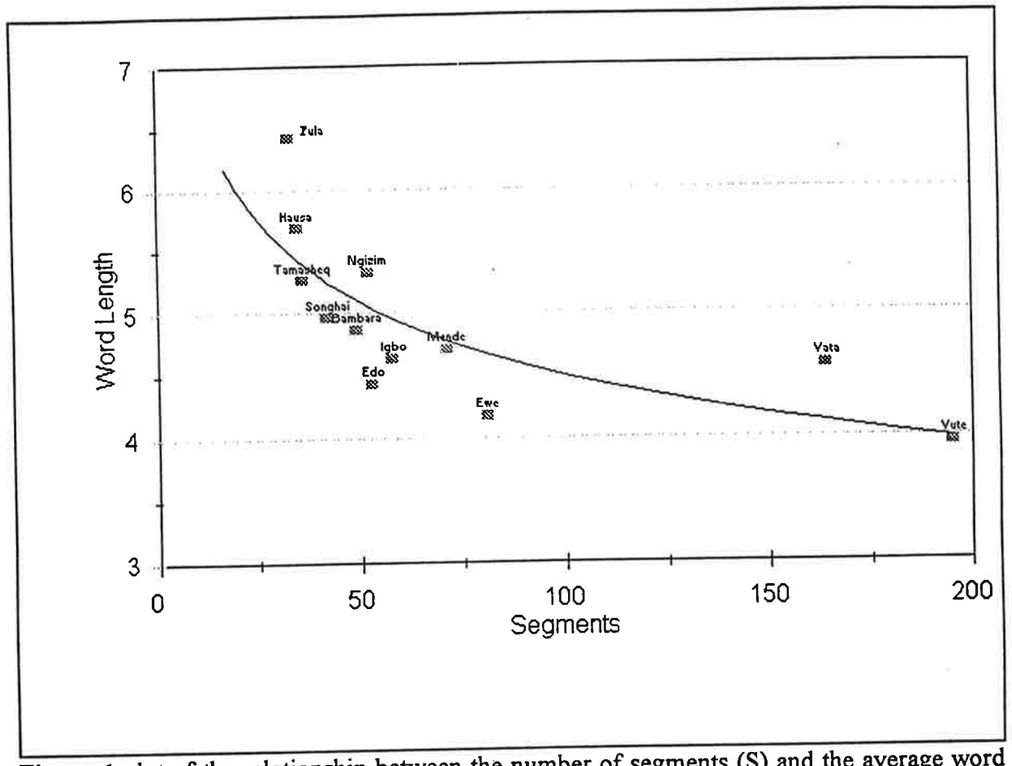


Figure 1 plot of the relationship between the number of segments (S) and the average word length (λ) for the twelve languages

Discussion: The Mechanisms of Language Adaptation

The observed relationship is a clear example of functional coevolution in language. Larger phonological inventories allow more economical coding of word-meanings, and the results show that languages with such inventories do optimise their codings accordingly. There is of course, a competing motivation, or all languages would evolve ever larger phonological inventories and very short words. The opposing motivation is perceptual: as the number of segments grows they become more phonetically similar, and so the probability of misperception and the difficulty of the hearer's task increase. Languages therefore reduce the number of their segments to facilitate decoding, and, as the results show, increase the length of their words to maintain contrasts in meaning.

The problem with results such as these lies in explaining how the adaptation of structure to function comes about. Zipf (1949), who discovered adaptive relationships such as that between the length of a word and its frequency, interpreted his results as the

outcome of the operations of a master-craftsman, who had fashioned language to be maximally convenient as a tool for communication. Even as metaphor, Zipf's explanation is misleading; languages are not designed by anyone. They are rather the products of an unintentional historical process.

I suggest that Zipf's creationist account of language design should be replaced with an evolutionary one. In linguistic performance, we observe a constant stream of minor variants on canonical word forms. If, due to least effort tendencies, learners are slightly biassed towards adopting the shorter of equifunctional forms, then over generations, word forms will always be reduced to the minimum length compatible with the preservation of meaning. Since, through coarticulation, segments colour those which precede them, this will lead to a piling up of phonological features earlier and earlier in words, and an increase in the total number of phonological contrasts.

Working against this, learners will fail to acquire distinctions between segments which are too similar phonetically. This causes phonological merger in languages and increases homonyny. Speakers compensate for homonymy with lexical expansions or innovations which are generally longer than the forms they replace. Thus reduction in phonological inventories leads to longer words.

Different languages thus represent different balances between these two adaptive pressures from speakers; to shorten word-forms and to merge segments which are too similar. In the full version of this paper I will also discuss historical reasons why some West African languages have found such different balances from others.

References

Köhler, R. (1986). *Zur Linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, R. (1987). Systems theoretical linguistics. *Theoretical Linguistics* **14**: 241-257.

Nettle, D. (1995). Segmental inventory size, word length, and communicative efficiency. *Linguistics* **33**: 359-367.

Nettle, D. (1996). *The Evolution of Linguistic Diversity.* Phd Thesis: University College London.

Zipf, G.K. (1949). *Human Behaviour and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley.

[Title] On Pitch Extraction
[Author] Shinya Matsunami
[Address] 445 Kaiolu St. #605, Honolulu HI 96815, U.S.A.
[Email] matsunam@hawaii.edu
[Affiliation] Linguistics, University of Hawaii (student)

[Summary] This article discusses the important aspects in analyzing language pitch patterns by using the pitch extracting instrument, including how to minimize the instrument-oriented and data-oriented errors, where to measure the pitch levels of the data, and how to identify various quantitative factors and qualitative factors, which must be controlled carefully.

[Words] controlling data, downtrends, errors, focus structure, Fo, measuring points, pitch extractor, quantitative (phonetic) factors, qualitative (phonological) factors, recording samples, statistical analysis, theoretical framework

[Topic area] 3. Methodological problems of linguistic measurement, model construction, sampling and test theory

---------- (Text) ----------

1. Introduction

Languages utilize pitch[1] to mark a syllable, a word, a phrase, and a sentence, etc., which is manifested as the pitch patterns of accent, tone, and intonation. In this article, I would like to discuss the important aspects in obtaining and analyzing phonetic pitch data by using the pitch extracting equipment. The present discussion assumes that the pitch extracting analyses are done under the theoretical framework in which the underlying phonological pattern is investigated from the surface phonetic pattern (e.g. Pierrehumbert 1980, Thorsen 1979, etc.).

The surface phonetic pitch forms of language is the manifestation of numerous phonetic and phonological factors interacting with one another. Following the practice of Pierrehumbert & Beckman (1988), I recognize the factors which are quantitatively analyzed as phonetic (such as the rate of declination), and the factors which are qualitatively analyzed as phonological (e.g. the presence/assignment of H and L tones).

The phonetic data are essential for the quantitative side of the analysis. Such substantial data are also indispensable to statistical analysis. Thus, the pitch extracting data must be very reliable. The material in the data corpus must be controlled in terms of the variables as well. However, unless the researcher is careful to control various causes of instrument-oriented or data-oriented errors, simple operation of a pitch extractor creates error-infested results.

2. Obtaining data

2.1. Using acoustic instruments

Pitch analysis instruments are mostly now all computerized. Some pitch extractors are installed in the hardware which are designed for that single purpose. However, most others are one of the functions of the sound analysis system, which has various other functions such as spectrograph, spectrum, and intensity analyses.[2]

As Hess (1982) and Rabiner et al. (1976) suggest, no pitch extractor is error-free. One can test the reliability of a pitch extractor easily by 1) extracting the pitch pattern of the same data many times in the same setting, 2) shifting the waveform horizontally on the time axis, 3)

[1]Technically, 'pitch' is the perceptual correlate of the fundamental frequency (Fo), whereas these two terms tend to be used interchangeably even in acoustic phonetics. The article also use them interchangeably.

[2]Some sound analysis software runs on a regular personal computer without any additional hardware. The personal computer in combination with the software should have a 16 bit sound capability for a professional quality analysis. (See Keller 1994.)

changing the vertical (frequency) or horizontal (time) scale ranges (either analysis or display ranges), or 4) using different algorithms (if the system has more than one). After all of these tests, a regular pitch extracting system will probably give inconsistent outcomes.

When the Fo pattern in the pitchgram differs from the more reliable pattern in the narrow-band spectrogram, it must be corrected. It is also important to compare the local Fo levels in the pitchgram with the frequency/time range of each wave in the expanded waveform.

Pitch extracting algorithms can be categorized into two major groups, i.e., the 'short term analysis' and 'time domain analysis' (see Hess 1982). The Fo detection in the short term analysis is performed based on the spectrum analysis, and the signal is analyzed by a small frame to average local pitch levels. Because of its analysis method, the short term analysis may create errors when the data contains a sudden Fo change.

In the time domain analysis, the Fo detection is performed based directly on the waveform. The algorithm tends to make errors when it reacts to the natural irregularities in waves too precisely, while its simpler calculation method provides a faster analysis.

The use of a filter (low-pass, high-pass, anti-aliasing, etc.) may often improve the reliability of the pitchgram especially when the recording quality of the data is not good. For example, time domain analysis tends to require the data to be filtered by a low-pass filter (to eliminate high frequencies) thereby reducing the possibility of picking up the higher frequencies rather than Fo. Sometimes, the pitch extractor picks up a higher harmonics rather than Fo, which can also be remedied by using the filter or by setting the 'frequency analysis range' properly.

If the speech analysis system has more than one type of pitch extracting algorithms, the researcher should try different types and find the one which best matches the data.

Adjusting the 'voice-voiceless threshold' level brings the pitchgram which best represents the voiced part . Proper settings of voice-voiceless threshold, as well as filters, often eliminates 'outliers' such as the very high frequency of a voiceless consonant.

When none of these options work, one should record the data again preferably in better recording conditions. Even if the same word is repeated three times in exactly the same manner, the analyses of each utterance may require different settings. Some utterance may lack the important information such as voicing of a vowel. Thus, one should record the same data more times than needed.

2.2. Recording equipment

Often it is possible to plug in a microphone (MIC) directly to the pitch extractor, then speech data can be directly recorded from an informant. This gives more reliable input to the computer than using an additional recording machine.

This method, however, has some limitations. First, it is often difficult to bring informants to the acoustic lab (where the instrument is). This is especially true when the researcher is doing field work or handling a great number of informants. Secondly, the acoustic lab itself may not be an anechoic chamber. An anechoic room should be used as a recording site as much as possible to avoid outside noises and echoes of the original signal.

Thus, it is often more convenient to use a separate recording machine (analog or digital), and the researcher can bring the recorded data to the lab and transfer it to the computer. It is important that the recording machine have a recording level control to obtain the optimum intensity level, and is not equipped with AGC (Automatic Gain Control), as this alters the intensity information of the raw data. (Most hand-held tape recorders have AGC which automatically functions at recording.)

Also, a high quality microphone that best-matches the recording machine or pitch extractor, should always be used. This makes a great difference in the recording quality.

3. Controlling the data for the pitch analysis

3.1. Intrinsic Fo effects of segments

Many phonetic experimental studies[3] report that pre-vocalic 'voiceless' consonants raise the Fo of the following vowel while 'voiced' consonants lower it. This Fo perturbation effects of pre-vocalic consonants especially affect the earlier part of the vowel. Thus, the vowel after a voiceless consonant tends to have a falling Fo contour whereas the vowel after a voiced consonant tends to have a rising contour. This Fo perturbation effect of the pre-vocalic consonant is widely attested as the cause of tonogenesis in the histories of many languages such as Chinese, Vietnamese, etc.

As for the post-vocalic consonants, some historical studies and a very limited number of phonetic experimental studies report that a [?] tends to raise the Fo of the preceding vowel, and a [h] tends to lower it.

According to numerous experimental studies, higher vowel such as [i] and [u] has higher intrinsic pitch than a lower vowel such as [a] when all other things are equal. However, there are few historical studies which propose the vowel intrinsic Fo as the cause of tonogenesis. It seems that the Fo perturbation effect of the pre-vocalic consonant appears to be more salient to the listener's perception compared with the vowel intrinsic Fo difference, while both phenomena are observable in the phonetic data. On the other hand, whether or not the Fo perturbation effect of the post-vocalic consonant are always present in the phonetic data appears to be a question. More studies should be conducted to find out whether it is a language-specific or position-specific (such as the utterance-final) phenomenon. Yet, in research, it is always necessary to control all consonants (pre- or post-vocalic) in the corpus material.

As Pierrehumbert (1980) suggests, the intrinsic Fo effects of segments are too great to ignore. To 'minimize' the interference of segmental intrinsic Fo effects, many phonetic studies use the data corpus only comprising the same vowels or vowels with similar heights and only voiced consonants.[4]

3.2. Downtrends

The downtrend is a general pitch lowering effect over an utterance. It is reported by many studies on different languages such as Dutch (t'Hart & Cohen 1973), English (Lieberman 1967, Cooper & Sorensen 1981, Liberman & Pierrehumbert 1984), Japanese (Fujisaki, Hirose, & Ohta 1979), Swedish (Bruce 1982), and some African tone languages (Clements 1979, Hombert 1974).

According to Beckman & Pierrehumbert (1986), the downtrend can be further divided into three different factors such as 'declination', 'final lowering', and 'catathesis.

Declination is a general and constant Fo downtrend over an utterance. It occurs without any trigger such as a certain tone sequence. Different causes of this type of downtrend have been suggested: the natural decrease of the subglottal air pressure (Lieberman 1967, Collier 1975), involvement of some preplanned behavior of the speaker (Cooper & Sorensen 1981), etc. Beckman & Pierrehumbert attribute it to the paralinguistic marking of the discourse structure by a speaker.

Final lowering is the pitch lowering phenomenon observed only at the end of a declarative utterance. Beckman & Pierrehumbert also attribute it to the paralinguistic marking of the discourse structure. The final sentence of the discourse shows the greatest final lowering while other medial sentences may show varying degrees of final lowering, depending on the meaning structure.[5]

[3] For the detailed reference in this section, please see Hombert, Ohala, & Ewan 1979. The full reference will be included in the text in the final version of this paper.

[4] The extreme example of this is the reiterant speech (see Nakatani & Schaffer 1978).

[5] According to Liberman & Pierrehumbert (1984), the final lowering affects a half the second from the end of a sentence (uttered in isolation) in English. So, in a long sentence, that duration may contain several syllables, but in a very short utterance like a word, the whole utterance may get the influence of final lowering.

Catathesis, another downtrend phenomenon, is triggered by certain (phonological) tone sequence such as that of an accent, which reduces the pitch level[6] of the succeeding material. This type of phenomena is reported in some studies as 'downstep' in some African tone languages (see Clements 1979) or analyzed with a sort of accent-reduction rule as suggest by McCawley (1968) in Japanese.

While Pierrehumbert & Beckman quantitatively analyzed declination, final lowering, and catathesis to some extent, much is still unknown about the behavior of these downtrend phenomena. However, a researcher must carefully take into account the presence/effect of these factors in the phonetic pitch analysis in relation to the underlying phonological form. (E.g. The two peaks with the same Hz readings in a pitchgram may not represent the same H tone phonologically.)

3.3. Focus structure

Many languages such as English (Liberman & Pierrehumbert 1984), Japanese (Pierrehumbert & Beckman 1988), Swedish (Bruce 1977), Dutch (Kruyt 1985), Danish (Thorsen 1980), Mandarin (Shih 1987), and Hausa (Inkelas, Leben, & Cobler 1986)[7] are found to possess the means to put focus on some word or phrase in a sentence by using a wider voice/pitch range over the focused sequence or (a) narrower/reduced pitch range(s) over the unfocused material in the rest of the sentence. Therefore, a H tone on a focused word often shows phonetically a much higher pitch peak than that of an unfocused word. The researcher should be aware of the presence of such focus structure in the language data under investigation.

4. Where to measure Fo in the data?

Studies are diverse in terms of the Fo measurement points in the data. It can be the average Hz value over a vowel (or a longer sequence) or one point in the pitchgram (such as the highest point) over a vowel (or a longer sequence), etc. Thorsen (1979, 1980) measured two or more points in a voiced stretch to obtain the stylized representation of Danish sentence pitch patterns. Pierrehumbert & Beckman (1988) measured the points in the pitchgram which are supposed to have some tones (H, L, etc.) according to their phonological model in English and Japanese sentences.

The present author is currently investigating the sentence pitch patterns of Mortlockese (a Micronesian language), in which the Fo levels are measured from the two points of each vowel. Based on the highest points of Fo, two halves of a vowel are determined to be a higher half and a lower half. In the higher half, the Fo value of the highest pitch peak is measured while that of the lowest pitch point is measured from the lower half. This method helped the analysis show the overall rising or declining stretches in a sentence very well, as the sentences in the language tend to show rising pitch toward the accent H tone, and declining pitch toward the end.

Generally, there are no established standards or guidelines for determining the measurement points. Depending on the goal, target language, available resources of the study, a consistent guideline has to be made by the researcher.

5. Looking for the phonological structure

After separating the phonetic factors as mentioned in the previous sections, the phonological structure will become much more visible.

To analyze the phonological structure, the researcher should be acquainted with different

[6] More precisely, it compresses the 'voice pitch range' of the speaker over the succeeding material.

[7] All these studies are quoted in Pierrehumbert & Beckman 1988.

phonological frameworks/theories which are most suitable to describe the language phenomena under investigation. One may need to modify other theories to make own model to describe certain language phenomena. It is helpful if the theory is equipped with the devises to factor out various phonetic and phonological phenomena as mentioned so far.

6. Statistical analysis

To identify and separate various phonetic and phonological factors, one needs to know how much the effect of some phonetic factor (e.g., the rate of declination) is, and how regular a phonological structure (e.g., the location of an accent H tone) is. To obtain generalizable qualitative trends from the results of a quantitative analysis, one should perform statistical analyses and obtain a certain level of significance (e.g. $P \geq .05$, $P \geq .01$, etc.). The statistical significance level also provides the inferential information.

In order to conduct a valid statistical analysis, one has to use well-controlled data in terms of segmental and suprasegmental structure. This is, however, not always easy as languages are often limited in their inventory.

Moreover, the use of statistics is not well-founded in linguistics while it is more widely used in experimental phonetics. The recent trend in the pitch pattern analyses of language, as I am discussing here, requires us to analyze both phonetic and phonological sides. If the study involves statistical analyses in more linguistics-oriented areas, more often the researcher faces the new situation in which no precedent exists. A researcher is then expected to make an extra effort to creatively apply statistics.

5. Conclusion

The pitch extracting analysis is important to investigate the suprasegmental system of a language. It involves preparing a well-controlled data corpus, recording high-quality samples, conducting the instrumental analysis with minimum errors, measuring the pitch levels from well-defined points, identifying various phonetic and phonological factors, and conducting all of these under a capable theoretical or methodological framework.

References (a synoptic version)

Beckman, Mary E. & Janet B. Pierrehumbert. 1986. "Intonational structure in Japanese and English." *Phonology Yearbook* 3: 255-309.

Bruce, Gösta. 1977. "Swedish word accents in sentence perspective." *Travaux de L'institut de Linguistique de Lund* 7.

Bruce, Gösta. 1982. "Textual aspects of prosody in Swedish." *Phonetica* 39: 274-287.

Clements, George N. 1979. "The description of terraced-level tone languages." *Language* 55.3: 536-558.

Collier, R. 1975. "Physiological correlates of intonation patterns." *JASA* 58: 249-55

Cooper, William E. & John M. Sorensen. 1981. *Fundamental frequency in sentence production*. Heidelberg: Springer.

Fujisaki, Hiroya, Keikichi Hirose, & Kazuhiko Ohta. 1979. "Acoustic features of the fundamental frequency contours of declarative sentences in Japanese." *Annual Bulletin, Research Institute of Logopedics and Phoniatrics* 13: 163-173.

t'Hart, J. and A. Cohen. 1973. Intonation by rule: a perceptual quest." *Journal of Phonetics* 1: 309-327.

Hess, Wolfgang J. 1982. "Algorithms and devices for pitch determination of speech signals." *Phonetica* 39: 219-240.

Hombert, Jean-Marie. 1974. "Universals of downdrift: Their phonetic basis and significance for a theory of tone." William R. Leben (ed.), *Studies in African linguistics*, suppl. 5. Department of Linguistics, UCLA.

Hombert, J. M., J. J. Ohala, and W. G. Ewan. 1979. "Phonetic explanations for the development of tones." *Language* 55: 1.37-58

Inkelas, S., W. R. Leben, & M. Cobler. 1986. "The phonology of intonation in Hausa." In S. Berman, J. Choe, and J. McDonough, eds., *Proceedings of the 16th Annual Meeting of NELS, GLSA*. University of Massachusetts, Amherst.

Keller, Eric. 1994. *Signalize User's Manual*. Lausanne, Switzerland: InfoSignal Inc.

Kruyt, J. G. 1985. *Accents from Speakers to Listeners*. Ph.D. Dissertation, University of Leiden.

Lieberman, P. 1967. *Intonation, Perception, and Language*. Cambridge, MA: MIT Press.

Liberman, M. & J. B. Pierrehumbert. 1984. "Intonational invariance under changes in pitch range and length." In M. Arnoff & R. T. Oehrle, eds., *Language Sound Structure: Studies in Phonology Presented to Morris Halle*, 157-233. Cambridge: MIT Press.

McCawley, J. D. 1968. *The Phonological Component of a Grammar of Japanese*. The Hague: Mouton.

Nakatani, Lloyd & Judith A. Schaffer. 1978. "Hearing 'words' without words: Prosodic cues for word perception." *JASA* 63.1: 234-245.

Pierrehumbert, Janet B. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. Dissertation, MIT. [Distributed by the Indiana Univ. Linguistics Club, Bloomington.]

Pierrehumbert, Janet B. & Mary E. Beckman. 1988. *Japanese Tone Structure*. Cambridge: MIT Press.

Rabiner, L. R., M. J. Cheng, A. E. Rosenberg, & C. A. McGonegal. 1976. "A comparative performance study of several pitch detection algorithms." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-24, No. 5.

Shih, C. L. 1987. *The phonetics of the Chinese tonal system*. Ms., AT&T Bell Laboratories, Murray Hill, New Jersey.

Thorsen, Nina. 1979. "Interpreting raw fundamental-frequency tracings of Danish." *Phonetica* 36: 57-78.

Thorsen, Nina. 1980. "Neutral stress, emphatic stress, and sentence intonation in Advanced Standard Copenhagen Danish." *Annual Report* 14.121-205. Institute of Phonetics, University of Copenhagen

A computer representation of articulatory gestures related to word phonotactics and speech sound classes

Antti Iivonen
University of Helsinki

Department of Phonetics
PL 35
Vironkatu 1
FIN-00014 HELSINGIN YLIOPISTO
FAX +358-9-191 8671
e-mail <AIIVONEN@Helsinki.Fi>

Project note.
Topic area: 2 application of methods: language comparison (key words: phonotactics, frequency of sounds and phonotactic patterns, articulatory movements, speech corpora)

Abstract:
A computer program simulating graphically the articulatory gestures of different languages, dialects and speech corpora is presented. The movements of the articulatory organs during continuous speech depend on the following language specific properties: phoneme paradigm, phonotactics, and the statistical frequency of the occurrence of phonotactical patterns. A graphical representation of articulatory movements is presented. The basis for the presentation is the classificatory consonant and vowel chart of the *International Phonetic Alphabet* simulating the articulatory space. Different statistical properties are indicated: the number of the occurrences of each sound type, the percentage of each sequence of two phonemes (or sound types), the percentages of voicing and the number of vowels (compared to that of consonants).

Because of the frequently occurring phonotactical combinations of a specific language, the speakers of that language repeat the articulatory movements corresponding to those combinations throughout their lives.

It is possible to illustrate the articulatory movements in a naturalistic way using the X-ray film (or similar) technique, but it is hardly possible to show the natural articulatory gestures in a speech corpus including **all** the possible phonotactical combinations of a language. A naturalistic articulatory description would include also coarticulatory and double articulatory phenomena. Instead of a naturalistic description, a representation based on IPA speech sound classification is suggested. A computer program indicating the articulatory movement on a chart resembling the combination of a IPA consonant and vowel charts can be applied (Fig. 1). Any phonetic transcription can be changed into its graphical IPA representation utilizing the ASCII code (cf. Wells 1987, 1989). Roughly speaking, the vertical position of these charts corresponds to the openness degree and the horizontal position to the frontness-backness dimension. Applying different language samples, the differences between **articulation bases** could be demonstrated.

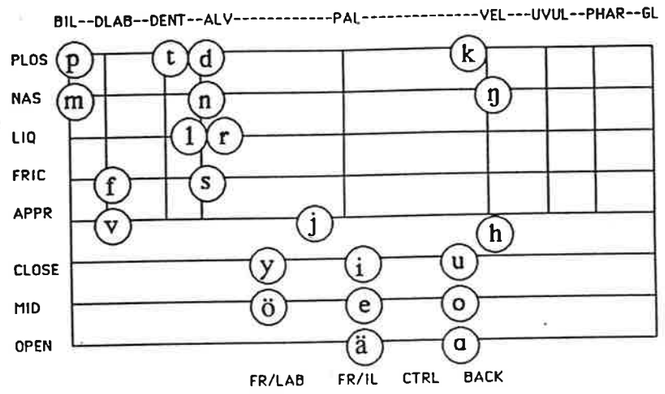


FIGURE 1.
The IPA chart used for the simulation of the articulatory space. The Finnish consonants and vowels are indicated on the chart.

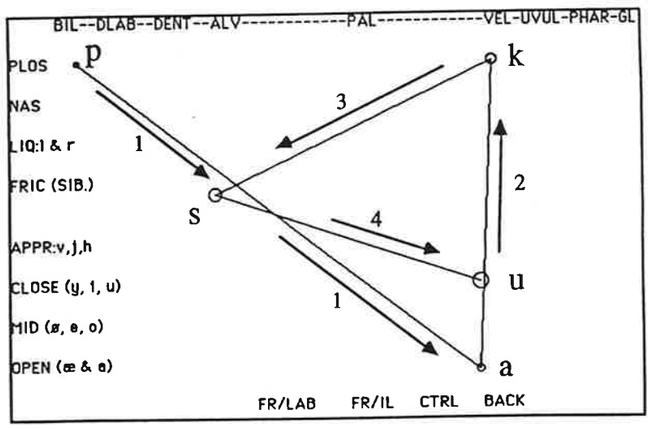


FIGURE 2.
The articulatory movement in the Finnish word *paksu* presentend in the classificatory articulatory space.

The program, written in the FutureBasicII language, shows the articulatory movements on the chart by means of the lines drawn between the points (Fig. 2; graphically x/y points on the pixel chart) corresponding to the adjacent speech sounds (allophones or phonemes). In addition, the **phonotactical position** of the sound is indicated by the size of the circle: the later the phonotactical position, the longer the radius of the circle. In Fig. 2, the articulatory movement in the Finnish word *paksu* 'thick' is illustrated. The lines combine the circles indicating the positions of *p*, *a*, *k*, *s*, and *u*. The size of the circles increases according to the phonotactical position of the phoneme. Furthermore, the percentages of the voiced sounds and vowels can be indicated. The percentages of the movements between two adjacent sounds (like that from *p* to *a* in *paksu*; cf. Fig. 2) can be expressed.

The general outlook of English (Fig. 3) is very different compared to Finnish (Fig. 4). The smaller circles illustrate the fact that English words are shorter than Finnish words. Because of the differences in phoneme paradigms, the graphical representation of Finnish and English have different architectures.

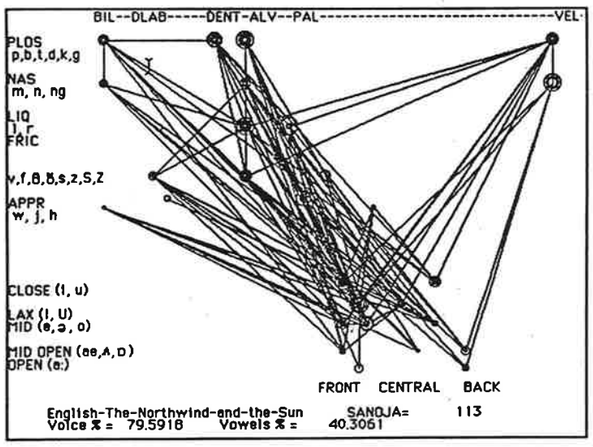


FIGURE 3.
The articulatory movements in the English text *The Northwind and the Sun* (The Principles of the International Phonetic Association 1949: 20). Number of words, the percentages of the voiced sounds and that of the vowels are indicated. The size of circles indicates the phonotactical position in the word structure: the bigger the circle the later the position of the sound in the word.

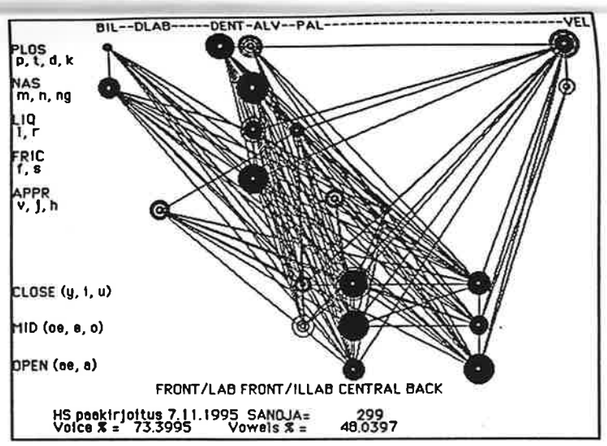


FIGURE 4.
The articulatory movements in the Finnish newspaper text (leading article in *Helsingin Sanomat* 7.11.1995.

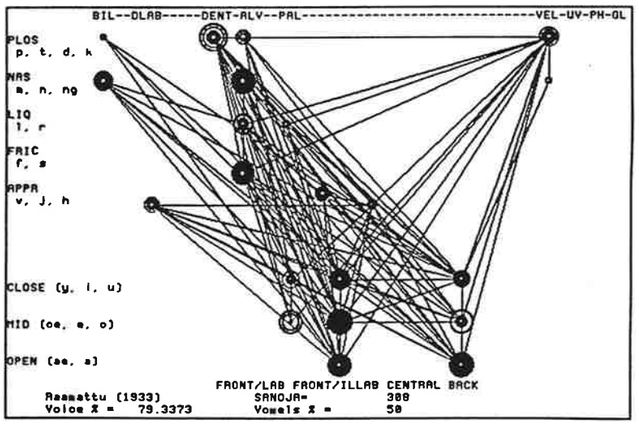


FIGURE 5.
The articulatory movements in the Finnish Bible (308 words from the beginning; translation in 1933).

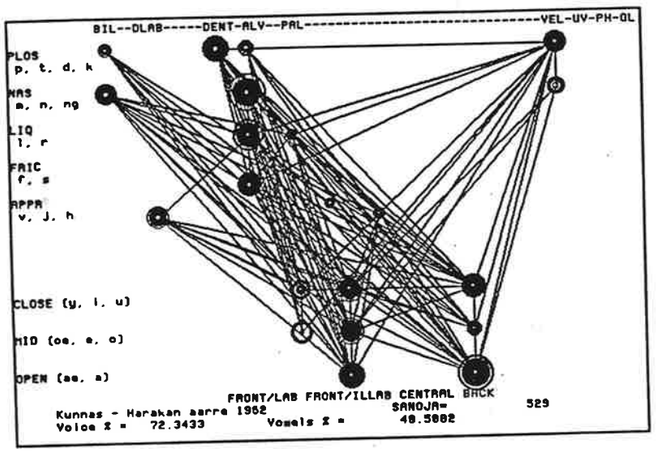


FIGURE 6.
The articulatory movements in the Finnish fairy tale *Harakan aarre* by Kirsi Kunnas.

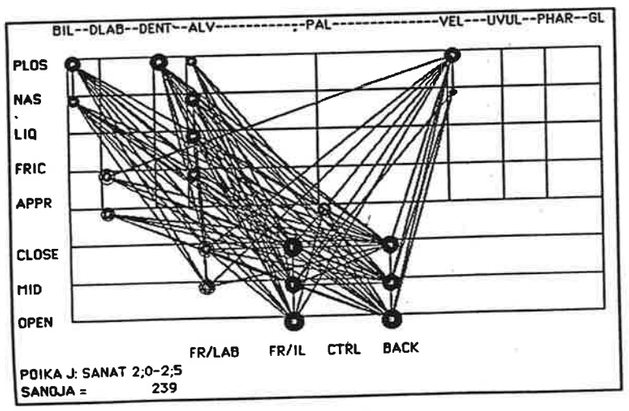


FIGURE 7.
The articulatory movements in 239 Finnish words produced by a child (age 2;0–2;5).

The Finnish illustrations (Fig. 4, 5, 6; partly 7) have more similar general outlooks. Observations concerning the relative positions of the phonemes can be made. In Finnish, /p/, /ŋ/, /h/, and /r/ seem to occur in relative early positions (the corresponding circles are small; cf. all Finnish examples). The child words (Fig. 7) are also much shorter than those of Standard Finnish and therefore the circles are much smaller. The general movement pattern resembles, however, the adult pattern already quite well.

The relative number of the occurrences of a phoneme can also be indicated utilizing the same IPA chart. Then the circle size corresponds to the relative occurrence of the phoneme (Fig. 8).

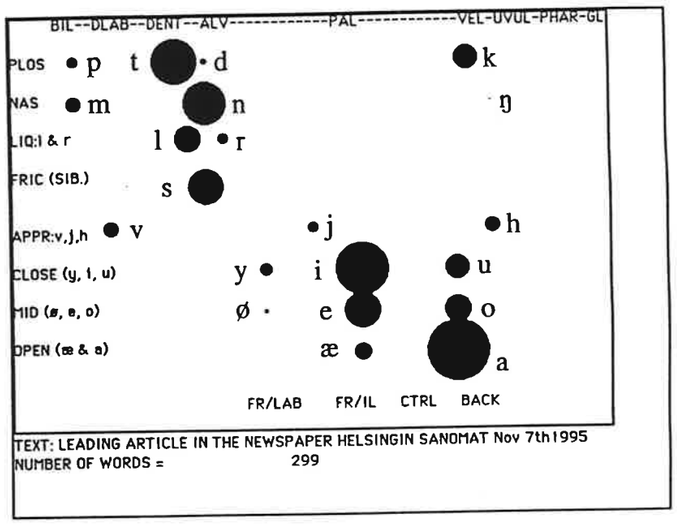


FIGURE 8.
The relative frequency of the phonemes in a Finnish text (leading article in *Helsingin Sanomat* 7.11.1995). The circle size corresponds now to the relative occurrence of the phoneme. Cf. Fig. 4.

BIBLIOGRAPHY

Wells, J.C. (1987) Computer-coded phonetic transcription. *Journal of the International Phonetic Association 17:2*, 94–114.
Wells, J.C. (1989) Computer-coded phonemic notation of individual languages of the European Community. *Journal of the International Phonetic Association 19:1*, 31–54.

Yuri Tambovtsev,
Novosibirsk Ped.university

The Distances between the Finno-Ugric Languages Derived on the Basis of Distribution of Consonantal Group Patterns in their Speech Chains

Every language has its sound pattern utilizing its phonemes with certain frequencies.Thus,it draws a unique sound pattern.We can construct the distances between languages basing on their unique sound patterns.The difference in the frequency of occurrence certain main phonological features is considered the distance.Though the seperate phonemes in some languages may differ drastically, we can find some universal phonological characteristics.Basing on the work of the active organ of speech(or place of articulation),the manner of articulation and the work of the vocal cords, we derive the eight basic features for consonants:1)labial; 2)front; 3)medio-lingual(or palatal); 4)back(or velar); 5)sonorant; 6)occlusive; 7)fricative;8)voiced

Table 1

Basic Phonological Feature Values in Finno-Ugric Languages for Consonants Derived by the Frequency of their Occurrence, %

Language/Feature	Lab.	Front	Palat.	Back	Son.	Occl.	Fric.	Voiced
1.Mansi(Sosv.)	13.55	30.09	6.79	10.64	34.76	17.00	9.31	0.00
2.Mansi(Kond.)	12.29	29.72	12.30	8.46	30.07	16.56	16.15	4.50
3.Hanty(Kaz.)	12.60	30.63	7.60	8.61	30.96	17.19	11.48	0.00
4.Hanty(East.)	10.45	30.81	5.19	13.53	21.82	24.20	13.96	8.06
5.Hungarian	10.17	34.72	3.71	9.30	22.61	22.58	12.71	12.15
6.Komi-Zyrjan	10.27	31.03	11.37	6.00	25.52	20.56	12.59	9.39
7.Mari(Gorn.)	9.99	33.90	6.06	7.92	24.61	16.35	16.91	9.43
8.Mari(Lugov.)	9.47	37.95	1.90	9.28	23.81	18.22	16.57	8.89
9.Mordov.(Erz.)	13.72	36.78	1.76	7.44	23.37	21.36	14.97	11.42
10.Veps	11.11	24.87	10.46	11.52	19.30	24.70	13.95	13.97
11.Vodian	11.95	33.62	2.68	7.66	20.71	21.93	13.26	8.50
12.Karel(Tihv.)	9.66	24.79	9.83	9.89	21.73	18.77	13.67	7.45
13.Karel(Livv.)	11.16	29.77	4.63	10.57	21.85	20.22	14.06	9.91
14.Karel(Ljud.)	8.09	32.00	1.33	9.34	17.45	20.56	12.30	10.40
15.Finnish	8.73	34.44	2.19	8.75	23.32	18.00	12.79	6.60
16.Saami	11.44	36.01	4.51	10.14	25.87	23.55	12.69	8.79
17.Udmurt	10.64	34.81	1.59	8.71	22.07	22.36	11.32	12.90

SEMANTICS

A semantic typology of Finnish verbs: Statistics for role frames

Pauli Saukkonen
Helsinki

The Department of Finnish and Saame, University of Oulu, hosted a big project on the "Semantic analysis of Finnish verbs" during the years 1988-95. The material was provided by the Dictionary of Contemporary Finnish (Nykysuomen sanakirja 1-6, Helsinki 1961, with ca. 200 000 entries). The analysis was performed by MA Jukka-Pekka Hammari. The coded material was then computationally processed by the collaborating partner, Research Institute for the Languages of Finland. After certain procedures the information will finally be added to the database of the Finnish lexicon.

All verbs have been analysed according to their nominal frame elements. The set of frame elements has been conceived of in a wide sense. Included are not only arguments, complements or obligatory actants (cf. Sue Atkins), but also other kinds of optimal elements. The semantic description is based on the semantic roles of the nominal frame elements, and the different meanings of one verb are thus accounted for in terms of role combinations (= role frames). The system will be illustrated with some examples below.

The analysed material consists of 383 865 instances of verbs (with clause examples covering 11 810 printed pages). The larger project also takes into account rough distinctions of inherent meanings in frame elements, but only the distinctive role frames have been taken into account for this paper.

The semantic roles in this kind of frame semantics are mostly equal with commonly used argument roles, but some additional roles are used as well (e.g. in the case of predicate complements). The main types can be seen in the following English examples.

INFLUENCER The bad condition on the roads	caused	RESULT accidents
PROCESSOR The car	is driving	WAY along the road
INFLUENCER-PROCESSOR (AGENT) The man	came	ORIGIN out of the house
EXISTENT The box	is	LOCALITY on the table
INFLUENCER-EXISTENT (AGENT) The man	is standing	
PROCESSOR I	met	OBJECT him
INFLUENCER-PROCESSOR (AGENT) I	made	RESULT a table

INFLUENCER-PROCESSOR (AGENT) She	makes	VARIABLE the table	RESULT long
INFLUENCER-PROCESSOR (AGENT) The woman	lengthens	VARIABLE the table	PROCESSOR with a board (INSTRUMENT)
INFLUENCER-PROCESSOR (AGENT) You	treat	PATIENT me	QUALITY well
INFLUENCER-PROCESSOR (AGENT) The man	pulled	PROCESSOR the car	ORIGIN out of the ditch PROCESSOR with a rope (INSTRUMENT)
INFLUENCER-PROCESSOR (AGENT) She	is holding	EXISTENT a candle	LOCALITY in her hand
INFLUENCER-PROCESSOR (AGENT) The man	gave	PROCESSOR the book	BENEFACTIVE to me

In my paper I will show some statistics for the role frames:
(1) what are the possible role frames,
(2) what are their frequencies in the lexicon,
(3) what are the different frame combinations occurring for one verb,
(4) what are their frequencies in the lexicon,
(5) what are the total frequencies for each role,
(6) what are the frequencies for each role for polysemous verbs,
(7) what are the frequencies of different frames where each role is occurring,
(8) what are the frequencies of verbs where each role is occurring?

As a result we will have a typology of Finnish clauses and distributions of verbal meanings. Finding out about these verbal meanings will also give us some answers to the following questions: how we construct reality, what kinds of frame elements of processes we need and what kind of a world view seems to be dominant in Finnish.

References

Hammari, Jukka-Pekka (1990). On the classification of verbs. In: Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja 30. Oulu.

Saukkonen, Pauli (1992). Lexical semantics and synergetics. In: Pauli Saukkonen (ed.), What is Language Synergetics? Acta Universitatis Ouluensis, Series B, Humaniora 16. Oulu.

Groups of English Verbal Synonyms

Ljalkova I.
Smolensk State
Pedagogical Institute
Russia, 214000 Smolensk,
Przhevalsky Str., 4
Phone: (081-22) - 37700

Topical paper
AREA: Quantitative analysis of groups of English verbal syno nyms.
Summary:
The study of systematic characteristics of synonymic groups of English verbal synonyms is given in this paper. The results were obtained by means of the method of correlational analysis (Pearson's criterion) on the total verbal body of Webster's Dictionary of Synonyms.

In Webster's Dictionary of Synonyms (Springfield, Mass., 1973) there are 543 verbal groups of synonyms. The number of synonyms in one group varies from 2 to 13. The average norm is from 4 to 8 members. There is no group of synonyms embracing more than 13 verbs. The deviation from the norm (6) is observed in 11% of cases. It is possible to say there is a certain "norm" and a certain "limit" to a group of synonyms.

I. Chronological Aspect

Chronological features characterize verbs from their origin in the Old English (OE), Middle English (ME) or New English (NE) periods. (See The Oxford English Dictionary on Historical Principles. Vol. 1-13, London, 1961). In a group of synonyms a pair of verbs can be of the same or different periods of the history of the English language. In the first case we speak of (A) synchronical, in the second - of (B) diachronical chronological correlations of the synonymous verbs.

A. Synchronical chronological correlations of verbs in a group can be of a different character:

1) verbs belong to OE period,

e.g.: bear - yield, be - live, cling - cleave, wane - ebb, behave - work;

2) verbs belong to ME period,

e.g.: admit - confess, ascend - mount, boast - vaunt, carry-transport, renounce - resign;

3) verbs belong to NE period,

e.g.: abduct-kidnap, actuate-motivate, ventilate-oxygenate, escort-chaperon.

B. Diachronical chronological correlations, in their turn, can be of the following character:

1) verbs of OE period correlate with ME verbs, e.g.: acknowledge (ME) - own (OE), climb (OE) - scale (ME), baptize (ME) - christen (OE), bear (OE) - convey (ME);

2) verbs of OE period correlate with NE verbs,

e.g.: play (OE) - personate (NE), compute (NE) - reckon (OE), lie (OE) - prevaricate (NE);

3) ME verbs correlate with NE verbs,

e.g.: demean (ME) - debase (NE), abdicate (NE) - demit (ME), comfort (ME) - accommodate (NE), sum (ME) - total (NE).

The correlational analysis of all types of chronological features in groups of synonyms proved:

1. Synchronical chronological characteristics correlate positively in their possible correlations (OE-OE, ME-ME and NE-NE).

2. Diachronical correlations within the groups of synonyms are not statistically relevant.

So, synchronical correlations within the groups of verbal synonyms are typical of the English language, while diachronical - are not.

II. Etymological Aspect

In a group of synonyms a pair of verbs can be of the same or different etymology (synonyms can be of the roman or german origin. In the first case we speak of (A) etymologically homogeneous, in the second - of (B) heterogeneous correlations between verbal synonyms.

A. Homogeneous etymological correlations are subdivided into two subtypes:

1) correlations between verbal synonyms of the german origin,

e.g. bring - take, char - singe, can - may, creep - craw, fish - angle;

2) correlations between verbal synonyms of the roman origin,

e.g. imbibe - assimilate, embrace - espouse, aqree - concur, value - cherish, cement - glue.

B. Heterogeneous etymological correlations of synonymous verbs in a

group can be examplified by the following pairs of synonyms, one of which is of the german, the other - of the roman origin,

e.g.: mistake (ON) - confound (OF f L), drive (Eng) - impel (L), outrage (Eng) - insult (L), perjure (OF, L) - forswear (Eng), pride (Eng) - pique (F).

The following results were obtained by the correlational analysis:

1. Etymologically homogeneous verbs correlate positively in groups of synonyms.

2. The correlation between etymologically heterogeneous verbs in groups of synonyms is negative.

So, etymological homogeneity of groups of verbal synonyms is typical of the English language, while heterogeneity - is not.

Conclusion: Groups of English synonymous verbs have a normal average of members, which cannot exceed 13, they have synchronical and etymologically homogeneous character.

The limited number of members of the group of synonyms makes to the process of language development. Still other synonyms being included into an extremely large group make other members change their meanings and fall out of the group. Small groups of synonyms, on the other hand, are open to new-comers untill they are ambiguous.

Borrowed and native words are included into groups of synonyms mainly in etymologically homogeneous and synchronical groups. Such pyramid reveals class - subclass relations where a group of etymologically homogeneous or synchronical synonyms (subclass) is included into a group of synonyms (class). Synonymy is the basis of the pyramid which is a constituent part of a language. Synonymy was not developed in the history of the English language but it existed from its early formation up - to- date.

References

1. Silnitsky G.et al. Correlations of verbal features of different language levels in English. Minsk, Navuka i Technica Publishers, 1990.

Dr.A.Ogoui (Czerniwzi, Ukraine)

BESTIMMUNG DES ANTONYMIEBEZIEHUNGEN ZWISCHEN DEN POLYSEMEN DEUTSCHEN WÖRTERN

Der Wortschatz ist als System organisiert, deshalb bilden nach E.Kurylowicz *(1962, 247f.)* zwei verschiedene Gebiete der Lexikologie - Polysemie und Synonymik das einheitliche Ganze, das noch mit einer anderen systemhaften Charakteristik - Antonymie in steter Verbindung steht. Das läßt sich erklären - Polysemie als Identität des Zeichens mit sinnverwandtem variablem Inhalt dient als Brücke zwischen der Synonymik (relative Identität des Inhalts mit variablen Zeichenformen) und der Antonymie (relative Kontradiktivität des Inhalts mit variablen Zeichenformen).

Die Verbundenheit dieser systemhaften Erscheinungsformen der Wortschatzorganisation soll auch im Lexikon (als Ausdrucksmittel der Sprachstruktur, das zu Lehrzwecken als Nachschlagebuch dienen kann) ausgedrückt werden. Leider bleibt diese aktuelle Frage weder theoretisch noch praktisch berücksichtigt - in den Bedeutungswörterbüchern werden Antonyme als wichtiges Systemcharakteristikum entweder vermieden , oder intuitiv asystematisch angeführt. Einige Lexikographen schlagen vor, die Antonyme in alphabetischer Reihenfolge dem zu beschreibenden Wort anzuordnen. Dadurch wird wenigstens der Systemcharakter der Sprache grob verletzt, geschweige vom Sprachgefühl des Lernenden. Die Antonyme gehören bestimmt zum Bedeutungswörterbuch, aber sie müssen dabei gemäß den existierenden Sprachbeziehungen angeführt werden.

Die Aufgabe der vorliegenden Arbeit ist das Anordnungskriterium der Antonyme zu finden. Dabei muß dieses Kriterium die existierenden Spracherscheinungen objektiv widerspiegeln sowie ziemlich formell sein, um die die maschinelle Sprachverarbeitung ermöglichen zu können.

Betrachten wir die Ausgangskategorien. Synonyme drücken "inhaltliche Übereinstimmung mehrer sprachlicher Zeichen bei verschiedener Lautform" (*Lewandowski 1984, 179*) aus. Man teilt sie in vollständige und partielle S. auf. Antonyme dagegen beziehen sich auf die "Verhältnisse gegensätzlicher Bedeutung bei etymologisch nicht verwandten Wörtern"(*Ulrich 1987, 23*). Dabei stützen sie sich auf das Vorhandensein qualitativer Merkmale, die sich quantitativ gradieren und/oder zum Gegnsatz führen lassen (vgl.*Lewandowski 1984, 70*). Sie werden in kontradiktorische (sich wechselseitig ausschließende: **Leben - Tod; männlich - weiblich),** kontrastive (gegenteilige: **groß - klein**) und konverse A. (umkehrbare: **gehen - kommen**) aufgeteilt. Die zahlreichen Diskussionen über die sog. vollständigen Synonyme und Antonyme brachten als Ergebnis die Schlußfolgerung, daß es keine vollständigen Synonyme und Antonyme gibt. Als Grund dazu dienen die Divergenzen in der Gebrauchssphäre, in den Eigenschaften der Kollokabilität und Kombinierbarkeit, in den stilistischen Schattierungen, in der Wortbildung usw. Dabei können die Wörter in bestimmten

Kontexten semantisch synonym und austauschbar sein, was als ihre relative oder partielle Synonymität betrachtet werden kann.

Maß der relativen Synonymität kann aufgrund der quantitativen Methode von S.Berezhan (1967) bestimmt werden. Demnach ist **M= 2C/n+k** gleich (wo **M** - das Maß der synonymischen Ähnlichkeit, **n** und **k** - die Anzahl der Bedeutungen bei diesen zwei synonymen Wörtern und **C** - die Anzahl der gemeinsamen Bedeutungen ist). Auf solche Weise entsteht die Basis für die formelle und objektive Bestimmung der Synonymie, wobei entsprechende Formel für die quantitative Berechnung der Antonymie noch fehlt.

Als bestimmter Ausweg kann die Anwendung der Formel von S.Berezhan (1967) für die Bestimmung der antonymen Beziehungen dienen. Betrachten wir das auf dem Beispiele der bekannten antonymen Bewertungsadjektive **gut** und **schlecht** (aufgrund der Wörterbücher DUDEN-UNIVERSAL (1997).

So ist **gut** in diesem Wörterbuch wie folgt beschrieben: 1.'bestimmten Ansprüchen, Zwecken genügend; von zufriedenstellender Qualität, ohne...Mängel' (*Ware; Nahrung);* 2.'angenehm, erfreulich; sich positiv einwirkend' (*Nachricht; Wetter);* 3.'gemessen; verhältnismäßig reichlich' (*Ernte, Appetit);* 4.tadellos, anständig; sittlich einwandfrei; auf eine religiös ethische Grundlage bezogen' (*Benehmen; Haus; Ruf);* 5.jemandem in engerer Beziehung zugetan, freundlich gesinnt' (*Benehmen; Ruf);* 6.'nicht für den alltäglichen Gebrauch bestimmt' (*Stube,Anzug);* 7.'leicht, mühelos geschehend' (*DUDEN-UNIVERSAL 1997, 644*; vgl. **schlecht** *DUDEN-UNIVERSAL 1997, 1326*). ***Schlecht*** ist: 1.'von geringer Qualität, viele Mängel aufweisend' (*Ware; Essen);* 2.'wenig schwach, unzulänglich' (*Gedächtnis; Gehalt; Esser);* 3.'ungünstig, nachteilig; nnicht glücklich...'(*Zeiten; Lage; Wetter);* 4.'Moralisch nicht einwandfrei,böse'; 5.'körperlich unwohl, übel'(s.*werden);* 6."schwerlich, kaum' (*s.sagen);* 7.(veralt) 'schlicht, einfach' (*s.reden).*

Erfassen wir die Sememe der antonymen Wörter schematisch. Dabei ist **gut:** 1.'Qualität; 2.günstig; 3.Quantität; 4.religiös moralistisch; 5.freundlich; 6.feierlich; 7.gutgelaunt, fröhlich im Geiste'. Ihm steht gegenüber **schlecht**: 1.'Qualität; 2.Quantität (unzulänglich); 3.ungünstig; 4.moralistisch negativ; körperlich unwohl; 6.schwerlich; 7.schlicht.

Wie daraus ersichtlich ist, sind vier Sememe dieser Antonyme kontradiktorische Gegenüberstellungen ('günstig - ungünstig' usw.) und ein Semem tritt als eine kontrastive Gegenüberstellung auf ('physisch unwohl - geistig wohl, fröhlich').

Gebrauchen wir diese Größen in der umgestalteten Formel von S.Berezhan, die jetzt als **A= 2C/n+k** (2) neuformuliert werden kann. Bei **C** = 4 ist **A** (als Maß der semantischen Antonymie dieser Wörter) 0,57 gleich; bei **C**=5 macht **A** eine statistisch relevante Größe 0,71 aus. Die neuinterpretierte Formel kann jetzt in den lexikographischen Beschreibungen verwendet werden, was bestimmte formelle Objektivität in dem zu schaffenden Bedeutungslernerwörterbuch (Czerniwzi,1997) gewährleistet.

The influence of morph-polysemy on morph-frequency

Andrea Krott

before 1/3/97:
Max Planck Institute for
Psycholinguistics
P.O. Box 310
6500 AH Nijmegen
The Netherlands
Tel.:++31/24/3615797
E-mail. akrott@mpi.nl

after 1/3/97:
Humboldt-Universität zu Berlin
Philosophische Fakultät II
Institut für deutsche Sprache und
Linguistik / Computerlinguistik
Jägerstr. 10/11
10099 Berlin-Mitte
Tel.: ++49/30/20192553
E-mail: akrott@compling.hu-berlin.de

A hypothesis about the dependency of morph-frequency on morph-polysemy will be discussed and tested on a German sample. It will be shown that the hypothesis seems to be correct for affixes and morphs which are used as verbs, nouns, or adjectives. In the case of morphs which are used as functional words the hypothesis cannot be confirmed.

topical paper

topic area: study in the field of synergetic linguistics

One goal of synergetic linguistics is to construct system-theoretical models of natural languages as a dynamic system. Its most important axiom is that the system is characterised by self-regulating and self-organising control mechanisms which help the system to adapt to all environmental requirements and to achieve an optimal steady state. This state can never be a final one because the environment and therefore also the requirements constantly change. Besides, this state always has to be a compromise because of different types of requirements which lead to cooperative and competitive processes inside the system. Like all systems the language system has a structure which is build up by entities and relations between these entities.

One of these relations is the frequency-polysemy relation. This relation is a general dependency which can be found in different areas where there exists some entities bearing one or more meanings and having a certain frequency, e.g. the dependency of word-frequency on word-polysemy. Already Zipf (1949:27ff.) has mentioned a dependency between frequency and polysemy of words. His idea was that frequency influences the number of meanings, not the other way round. He assumes that this influence is based on an economy principle: The actually used set of words can be kept small by using frequent words to express new meanings (cf. Zipf 1949:67). The models proposed by Köhler (1984), Hammerl (1991), and Gieseking (1993) treat the influence of frequency on polysemy as an indirect one: Frequency has an influence on word length which has an influence on polysemy. On the other hand is a word frequency dependent on the number of texts where it appears and this number of texts is dependent on polysemy (cf. e.g. Köhler 1986:74).

Words are not the only kind of entities in language which can have meaning. On the one hand there are smaller parts, the morphemes, and on the other hand there are combinations of words, e.g. phrases and sentences.

Directing our attention to morphs we assume that the more meanings a morph has the more often it is used as a part of a word, or - in a shorter way -

the frequency of a morph is dependent on its polysemic potential

In analogy to Köhler's approach we can formulate the frequency-polysemy-hypothesis as the following mathematical function:

$$\frac{y'}{y} = \frac{G}{x}$$

with y as the frequency of a morph, x as its number of meanings, and G as a proportional factor. This differential equation has the solution

$$y = C \cdot x^G$$

To test the hypothesis we have to find a method to count the number of meanings and the number of occurrences of a morph in a language. That has been done for the German language and with the help of dictionaries.
To get a list of morphs and a list of words where these morphs appear, the CELEX lexical database was used. In this database the entry for every word is split up into its morphemes. The frequency of a morph can be easily measured as the number of words which it is a constituent of[1].
Determining the polysemy of a morph is a much more difficult task. There are no morpheme-dictionaries where morpheme meanings are written down, except for affix-dictionaries.
The meanings of affixes in our morph list were counted in the following way: There is a morpheme register in 'Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache' by Kühnhold and Prell where affix meanings are notated as indices. The number of different indices has been taken as the number of meanings.
In the case of the remaining morphs we used a trick. All these morphs can form a word by themselves. So we have to determine the primary wordclass of these words and look them up in a dictionary. To do so the morphological analyses of the CELEX database were used, because every constituent in these analyses is annotated with its basic word class. In the case of homographical morphs we had to sum up the meanings of both words.
The word meanings were taken from a dictionary which is based on the 'Wahrig - Deutsches Wörterbuch'. This dictionary was built in the project 'language synergetics' at the Ruhr-Universität Bochum and the University of Trier. The dictionary contains 96,601 lemmas, but only 10,487 lemmas have been expanded by meaning hierarchies taken from the 'Handwörterbuch der deutschen Gegenwartssprache' which makes it possible to count the number of meanings of a lemma.
For an empirical investigation of the above hypothesis it has to be proved whether the equation describes the empirical data. To do so the equation was linearised:

$$\ln y = \ln C + G \cdot \ln x$$

To make this equation more readable we can also write

$$\text{L-FREQU} = \ln C + G \cdot \text{L-POLYSEMY}$$

Linearisation allows us to carry out a linear regression, which is a more reliable method than a non-linear regression because a non-linear regression can stop at a local minimum so that a global solution cannot be achieved.
To improve the goodness of fit the data was weighed. That has been done by weighing every x/$\bar{y}$-pair by the number of values of which $\bar{y}$ is a mean of. By doing so, it can be made sure that runaways do not disturb the result and that the resulting curve approaches the largest part of the data by neglecting parts with less data density.

[1] Morph-frequency can also be measured in text. It does not make a big difference because text frequency and word frequency of morphs are positively correlated (cf. Krott 1997)

The frequency distribution of the polysemy-values in Fig. 1 shows that weighing the data is an important issue for our investigation. In the sample 95% of all morphs have at a maximum 8 meanings and 8 is just 20.5% of the possible 39 meanings. Therefore it is more important that the resulting curve approaches the mean frequency of morphs with a polysemy-value of e.g. 2 than that it approaches the mean frequency of the only one morph with 39 meanings.

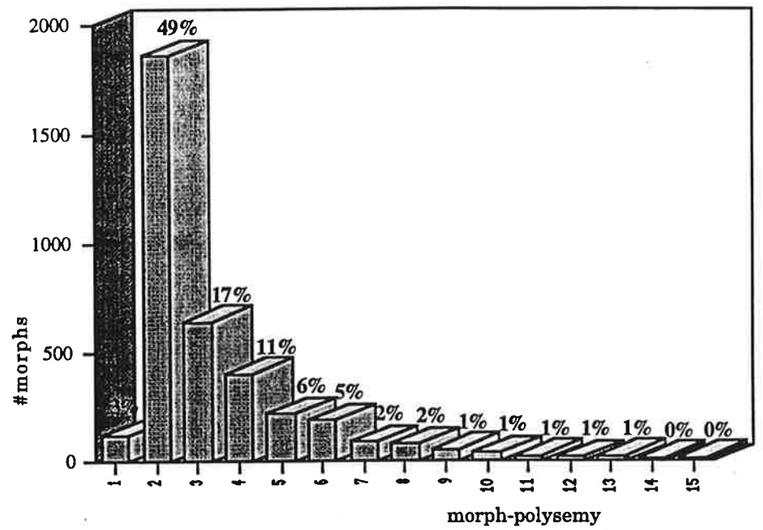


Fig. 1: Frequency distribution of the morph-polysemy (part)

C	G	R^2	Sign T
3.22460	1.15072	0.65107	0.0000

Table 1: regression of the function $\text{L-FREQU} = \ln C + G \cdot \text{L-POLYSEMY}$

The result of the regression analysis is shown in table 1. The value of the determination coefficient R^2 shows that the fit is fairly good. But the significance niveau t lies under 0.00005. T is not very reliable in the case of big sample sizes like this one, but it can be interpreted as a hint that the variables of the linearised model are linearly related. This assumption is also supported by the optical impression of fig. 2, which shows the sample data and the non-linear function estimated on the base of the sample. The curve follows the data especially in the part of small polysemy-values presenting the largest part of the data. The spreading of the data points increases with rising polysemy-values.
It is conspicuous that morphs with one meaning are exceptionally frequent. This relatively large value can be put down to the fact that there are a lot of affixes in this morph-group which have just one meaning, but which are very frequent. These were or still are productive affixes (e.g. -s: 1815 occurrences; -ier: 915; -keit: 801). Their productiveness may be an effect of their transparency, i.e. people like to build new words with the help of affixes which are unambiguous so that the meaning of the new word is easy to understand.
If we assume that the polysemy-frequency-relation exists, we have to find a reason why R^2 is only 0.65. A possible explanation for this is the lack of data homogenity due to the different types of morphs. We have already seen that morphs with one meaning are mostly affixes and that they seem to behave in another way. It is also possible that the

measurement of the morph-polysemy is insufficient. The unusual way of measuring the number of meanings on the base of word meanings is also a reasonable source of error. But it is difficult - perhaps even impossible - to find another method.

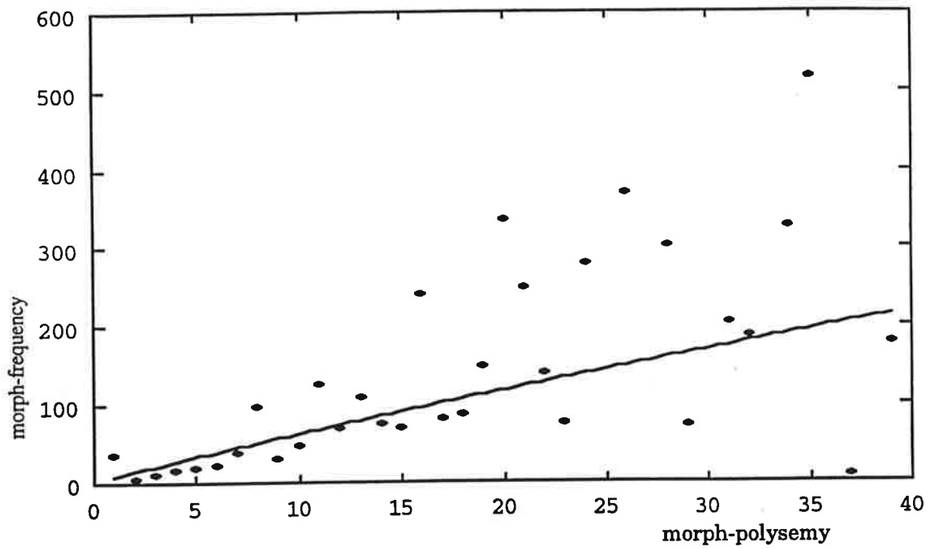


Fig. 2: The dependence of morph-frequency on morph-polysemy

We should not forget that determining meanings of affixes and of those morphs which are used as function words is problematic and that these types of meanings are totally different from those of nouns, verbs, or adjectives.
Because of all these reasons morph types were examined separately. The remaining morphs can be divided into classes by considering the annotations of different word classes in the CELEX database. Table 2 shows the results for all classes, except for numerals, determiners, and interjections whose numbers did not allow a regression analysis.

morphtype	A	B	R^2	Sign T
noun	4.0029	0.98843	0.92105	0.0000
adjective	3.75433	0.92173	0.79888	0.0000
verb	4.03457	1.01888	0.95774	0.0000
pronoun	7.60899	-0.1066	0.00959	0.7621
adverb	9.15024	0.21087	0.02513	0.5726
preposition	10.8281	0.82224	0.28052	0.0238
conjunction	1.79472	-0.0630	0.00935	0.7904
affix	45.2013	0.80148	0.67289	0.0000

Table 2: regression of the function $\text{L-FREQU} = \ln C + G \cdot \text{L-POLYSEMY}$ - different morph types

Because of the high values of R^2 the hypothesis can be accepted in the case of nouns, adjectives, and verbs. Affixes build a class for themselves. This can also be seen from Figures 3 to 5. In the case of pronouns, prepositions, conjunctions, and adverbs the hypothesis cannot be accepted because of the value for R^2.
These results lead to the assumption that morphs which are used as function words worsen the overall result. Maybe we should not speak of meanings in those cases at all, at

least not in the sense we understand meaning in the case of nouns, verbs, and adjectives. We cannot solve this problem here, but we have to emphasise that there should be a reconsideration of what the meanings or functions of such morphs really are and if there should be a difference made between morphs which are used as function words and morphs which are used as nouns, verbs, or adjectives. Maybe they cannot be treated in the same way.

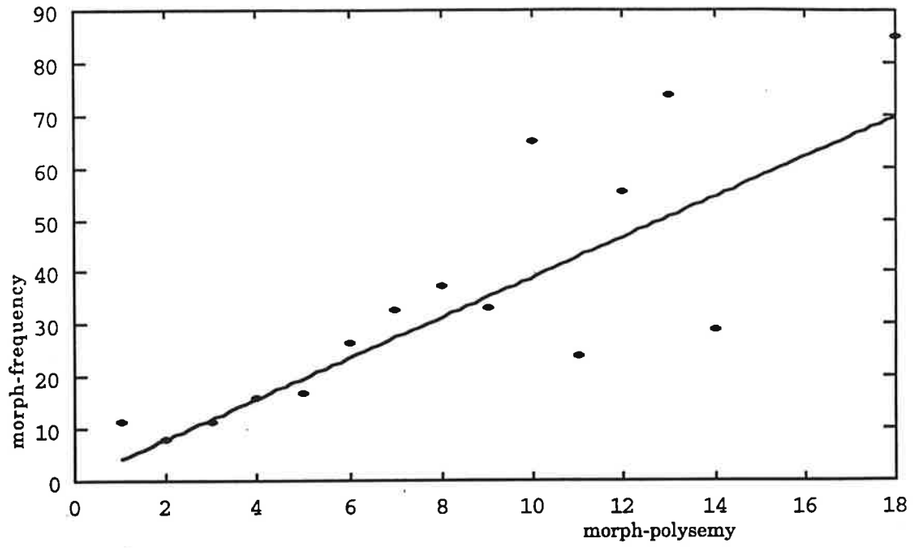


Fig. 3: The dependence of morph-frequency on morph-polysemy (nouns)

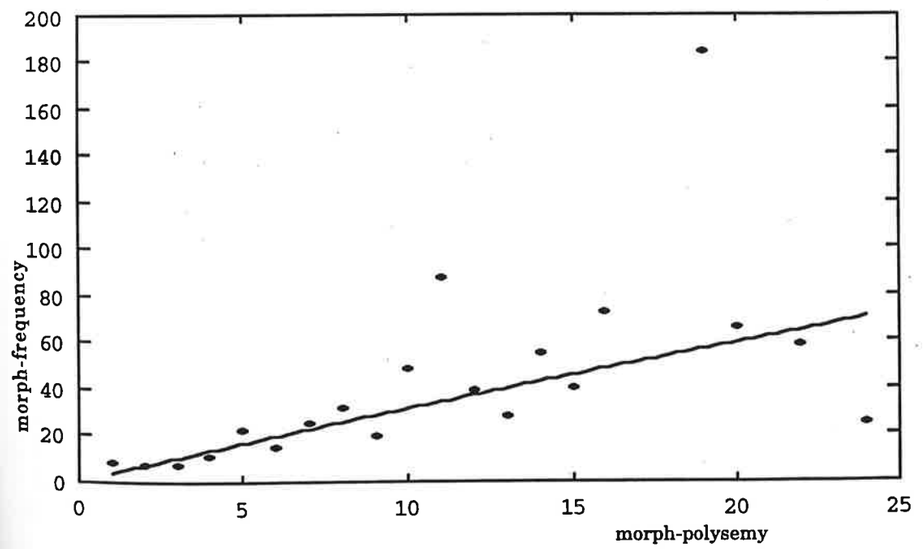


Fig. 4: The dependence of morph-frequency on morph-polysemy (adjectives)

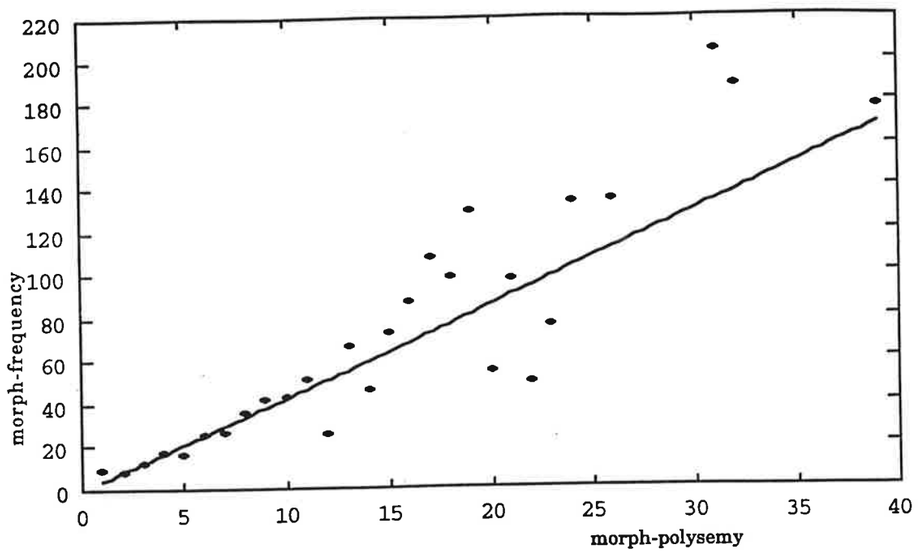


Fig. 5: The dependence of morph-frequency on morph-polysemy (verbs)

In the case of affixes, the results of the regression analysis support the hypothesis about the dependency of the morph frequency on the morph polysemy. The t-test shows that after linearisation the dependency is a linear one. The t-test is more significant in this case than in the case of the whole group of morphs because the set of 183 different affixes is relatively small and the t-test is more meaningful for a small number of data. Fig. 6 shows the data and the resulting curve for affixes. A morph polysemy over 7 increases the spreading of the data points, but this is just 10% of the whole sample.

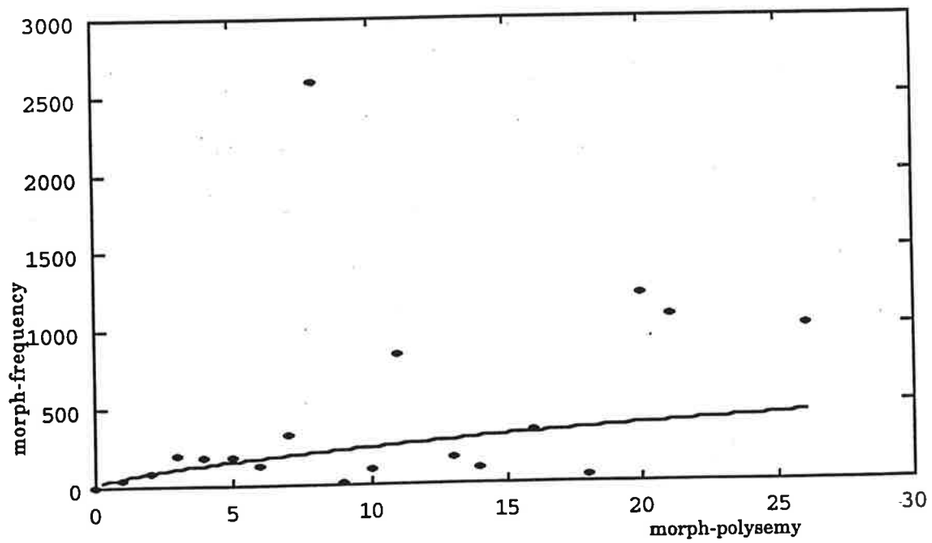


Fig. 6: The dependence of morph-frequency on morph-polysemy (affixes)

As a conclusion we can say that the existence of the dependency of morph-frequency on morph-polysemy has been confirmed, at least for affixes and morphs which function as

nouns, verbs, and adjectives. In the case of function words further investigations are necessary.
But after all another small part of the language system has been illuminated, which is another step on the way to a language theory able to explain language phenomena.

References

Baayen, H., R. Piepenbrock u. H. van Rijn (1993): The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

Gieseking, Kathrin (1993): Synergetische Aspekte von Struktur und Dynamik der Englischen Lexik. Unpublished masters thesis.

Hammerl, Rolf (1991): Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells. Trier.

Köhler, Reinhard (1986): Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. Bochum. (Quantitative linguistics; 31)

Krott, Andrea (1997): Ein funktionalanalytisches Modell der Wortbildung. (to appear)

Kühnhold, I. u. H.-P. Prell (1984): Deutsche Wortbildung, Typen und Tendenzen in der Gegenwartssprache. Morphem- und Sachregister zu Band 1-3. Düsseldorf. (Sprache der Gegenwart; 62).

Wahrig, G. (1985): Deutsches Wörterbuch. München.

Zipf, G. K. (1949): Human Behavior and the Principle of Least Effort. Reading, Mass.

RUSSIAN WORDS "RUSSKIY" AND "SOVETSKIY": SEMANTICS OF NOUN AS AN OBJECT OF QUANTITATIVE ANALYSIS OF ITS ASSICIATIVE FIELD

Vladimir A. Dolinsky, Ph.D.

V. Dolinsky, Moscow State Linguistic University. Department of Applied and Experimental Linguistics. Docent. 129345, Moscow, Ostashkovskaya, 9-2-98. Russia. Tel.: (095) 475-8384. E-mail: nalimov @ Nalimov.home.bio.msu.ru

What Russian word "RUSSIAN" means for Russians? Free associations obtained both from Moscow students in 1990th, from Moscow children in 1980th, and from adults (mixed group) in 1970th, are regarded in quantitative measurement. Also test data for "SOVETSKIY" are presented.

PROJECT NOTE.

Applications of methods from quantitative linguistics to problems of psycho- and sociolinguistics, philosophy and culture. Methodological problems of sampling and test theory.

Semantical analysis of Russian words "RUSSKIY" and "SOVETSKIY" is based on distribution of responses to these words as stimuli in word association tests.

Three massives of this collection have been obtained from subjects representing different linguistic communities:

1. Associative fields from experiment conducted by author with Moscow students in 1991-1994 (Massive MSA). Number of subjects is 1010.

2. Associative fields from experiment conducted by author with russian children 6-7 years old in 1987-1988 (Massive D). Number of subjects is 200.

3. List of associations obtained by A.Leontyev from 16--50 years old speakers in 1977 (Massive L). Number of subjects is near 250 (Data are partly - without tailes - presented on tables 1 and 2).

Each of informants gave only one response to the each word-stimulus. Mother tongue of all subjects is Russian.

Collection of associations will be of interest to the researchers in the field of verbal learning and verbal behavior. This research provides comparisons among these different sets of associations that point up consistent differences among groups of subjects residing on the same region (space) but not on the same cultural-linguistic field (time and age).

Changes in the meaning of stimuli over time play some part in the changes observed, for instance changes in frequency of responses (such as "Jew", "Slav", "German", "Georgian", "Chechen", "Chukcha", "Tatar", "Frenchman", etc.). Of special interest is the possibility to monitor the dynamics of change in the associative potential of words in relation to the age of respondents.

The possibility to analyse semantic characteristics of lexical units using data of psycholinguistic experiments by the methods of quantitative linguistics without ad hoc logical hypotheses is of great interest too.

Table One

ASSOCIATIVE FIELD "R U S S K I Y"

Massive L	Massive D	Massive MSA
S = 232	S = 200	S = 1010; N = 923
40 yazyk	79 soldat	147 yazyk
36 chelovek	63 chelovek	139 chelovek
12 narod	7 narod	33 yevrey
10 kazakh	6 flag	29 muzhik*
8 n'em'ec	5 Lenin*	23 narod
7 les	zv'ezda*	17 kharakter
6 n'em'eckiy	4 d'ad'a*	16 rodnoy
sovetskiy*	3 dom*	12 nazionalnost'
5 yevrey	tank*	11 Ivan
inostran'ec	2 zeml'a	nash
nazionalnost'	mal'chik	10 patriot*
4 rusyi	mashina*	9 slav'anin*
ukrain'ec	samol'ot*...	ya
francuzskiy		8 Rossiya*
kharakter		dukh*
3 velikiy		n'em'ec
nauka*		pasport*
svoy		svoy
slovar'		7 nerusskiy*
francuz		uzkiy*
2 Ivan		6 velikiy

drug		gordost'*
inostrannyi		dom
kitayec*		durak*
pisatel'		dusha*
poet		n'em'eckiy
rodnoy	5	gruzin
soldat		pisatel'
khoroshyi...	4	blondin
		kvas*
		prostoy*
		svetlyi
		chukcha*...

Table Two

ASSOCIATIVE FIELD "S O V E T S K I Y"

Massive L	Massive D	Massive MSA
S = 228	S = 200	S = 1010; N = 902
69 chelovek	87 soldat	137 chelovek
20 Soyuz	28 chelovek	106 pasport
13 nash	13 zv'ezda*	85 Soyuz
11 narod	12 Soyuz	43 grazhdanin
7 pasport	9 flag	25 sovok*
6 antisovetskiy	7 Lenin*	21 narod
5 inostrannyi	3 Moskva	20 stroy*
4 grazhdanin	drug	18 obraz zhizni*

zakon	rodina	16 krasnyi
socialisticheskiy	strana	15 plokhoy*
3 krasnyi	2 Kreml'*	flag
rodnoy	boyec	12 rayon
khoroshyi*	dom*	9 nash
2 gorod*	voin	8 diplomat*
gumannyi*	gerb	7 sovkovyi*
korabl'*	narod	6 voin
luchshyi	serp i molot*...	otlichnyi
otvetstvennyi*		russkiy
otlichnyi		5 proshlyi
patriot		soldat
patriotizm*		gerb
russkiy		4 gimn*
sovetskiy		durak*
uchonyi...		zakon
		patriot...

---

* - this sign markes responses with frequency zero in both other massives;

S (N) - number of subjects (responses)

The concept, i.e. the generalised image of a word, is made up of several Gestalten that vary in different languages as they vary the world pictures of their speakers just like pictures in children's kaleidoscope.

Analysis of associative fields serves as a means of revealing objective semantic characteristics of linguistic

units and is indispensable source of information on language -cultural archetypes.

"The method of dividing the field of thought by means of language variability (diversity) has been little tested as yet, but it does not become less possible or important because of it. However rich and fruitful a language might be, it is never possible to imagine the real sense, the sum total of all integrated characteristics of a word, denoting a non-physical object, as the definite and final value" (W.von Humboldt).

The main linguistic unit, a word, is regarded as embodied in its associative field - an image of unclosed, orded in hierarchy, specific multitude of other words connected with given one by associations inherent to subject and cultural-lingual society (group).

The associative field of a word serves as a key to revealing its sense. To understand a word means to set a weight function assigning different values to various sections of the associative field, or ranging different associations according to their force, stability, frequency, etc. Recurring connections between words are reproduced in cognitive and communicative processes, thereby fixing themselves into meanings in language and culture.

Dolinsky V.A. 1993. A model for word association // European Mathematical Psychology Group, 24-th Annual Meeting. Moscow.

Dolinsky, V.A. 1994. Moscow Students' Word Associations

// QUALICO'94. Moscow, pp.66-68.

Dolinsky, V.A. 1995. Kulturnoyazykovyye arkhetipy v associaciyakh // Ethnicheskoye i yazykovoye samosoznaniye. Matherialy konferenzii. Moscow, pp.38-40.

Dolinsky, V.A. 1996. Russian VERA and English BELIEF: Semantics of noun as an Object of Associative Analysis // Sociolinguistic problems in different regions of the world. Moscow, pp.152-155.

Cherneyko, L.O.; Dolinsky, V.A. 1996. The noun SUD'BA as an Object of Conceptual and Associative Analysis // Moscow State University Bulletin. Ser.9: Philology, N 6, pp.20-41.

PLENUM II

On the Probability of Probabilities

Some notes to stochastic approach to the Zipf formula

Jan Králík

Summary

The paper shows one of the possible ways how to deduce the Zipf formula by means of stochastic construction. The probabilistic interpretation of rank and a simple distribution function for measure of usage of words are suggested. The issue takes the form of Mandelbrot's correction of Zipf.

Words

probability, word frequency, empiric formula, distribution, stochastisation, Zipf

Topical paper

topic area 2 / 3

Address

RNDr. Jan Králík, CSc.,
Ústav pro jazyk český
Akademie věd České republiky
Letenská 4
118 51 **Praha** 1 - Malá Strana
Telephone +42 / 2 / 245 11 229 * 342
Fax: + 42 / 2 / 53 62 12
since March 1997: telephone +420 / 2 / 573 20 942 * 342
since March 1997: fax: + 420 / 2 / 53 62 12

On the Probability of Probabilities

Some notes to stochastic approach to the Zipf formula

/preliminary version, to be precised in details and conclusion/

Jan Králík

The empiric tendency as expressed by the famous Zipf formula was often referred to as a law, and, at the same time, it used to be criticized by a statement that it does not possess the character of a law, but of an empiric formula only.

First algebraic expressions, as published by Estoup (1916) and Condon (1928), has been based on the optical similarity between the curve describing the decrease of word frequencies and hyperbolic function. Because the conception of rank of decreasing frequencies does not represent any real variable, but a mere practical issue of ordering, main authors of quantitative linguistics (as, e. g., Herdan 1960) refused to accept the rank-frequency curve among other statistical laws. Such strict point of view excluded any consideration of the Zipf formula from the probabilistic point of view.

In this paper I should like show one hidden possibility of the purely probabilistic reading of the Zipf formula. Or, following the sequence of considerations, I should like show one possibility of the probabilistic construction of this formula, so that it could be involved among other statistical - and may be also natural - laws.

*

The simple re-ordering of any items according to the decrease of their frequencies is far away from usual statistical investigation of natural events. It is much more close to journalistic presentation of statistical data. In fact, one hidden interpretation still remains.

The re-ordered sequence of items helps show, which probability of occurrence is more probable and which one is less

probable. In other words, the re-ordered sequence of items represents an expression of how their probabilities of occurrence are distributed. In this, the well known regularities are found: on the level of words, e.g., the majority of different words possess the frequency 1 and/or 2 in any text, which refers to the lowest probabilities. And, on the other hand, we never face more than one word with the first, second etc. - highest frequency.

Generally: the minimum of words possess the highest frequencies referring to highest probabilities, and great deal of words possess the lowest frequencies referring to the lowest probabilities. The space between these extremes is structured sequentially. The un-regularities and inversions are much more exceptional, than they were rule.

The described tendencies occur in nearly every observation and their character does not change. The more: general character of the observed tendencies, whether it already was, or was not expressed by any empiric formula, shows that there should exist some more general and more commonly applicable distribution of probabilities not only for lexicon, but also for some other fields, which are results of human activities.

The following consideration about the distribution of such probability of probabilities will lead directly to an interesting stochastization of the Zipf formula. To make the explanation more clear, I shall substitute the term *"probability of a word"* by the *"measure of usage of a word"*.

The measure of usage of a word can be easily acknowledged as a random variable. If no transformations are being used, the measure of usage of a word takes values between 0 and 1. The distribution function of such variable Φ is, as usually, expressed by

$$F(x) = P(\Phi < x) \qquad (1)$$

The supplement of the distribution function in the point x

is then:

$$N\ (\ x\ ) = 1 - F\ (\ x\ ) = P\ (\ \Phi \geq x\ ) \qquad (2)$$

$N(x)$ expresses the probability that the measure of usage of a word is equal or higher than x.

*

This probability has its real representation in every frequency dictionary or frequency list by the ratio:

$$N\ (\ x\ ) \approx \frac{\textit{number of words with the measure of usage equal or greater than } x}{\textit{number of (different) words}} = \frac{r_x}{V} \qquad (3)$$

First, let us consider the beginning of the frequency list. For any x, the number of words with the measure of usage which is not lower than x, is equal to the rank number of the word, the measure of usage of which is not lower than x. Then, subsequently, it can be written

$$N\ (\ x\ ) = 1 - F\ (\ x\ ) = r_x\ /\ V \qquad (4)$$

where r_x is the rank of the word, the measure of usage of which is equal or greater than x.

Second, let us consider the ending of the frequency list. Following the reversed direction within the word list, there are long intervals of words, the measure of usage of which is equal to *1*, *2*, *3*, etc. In such situation, a non-statistical ordering of words is used by application of alphabetic order. Because of this, from the point of view of probability, the rank is random within such interval. In fact, not only theoretically, even within such intervals the measures of usage of words differ. In the case of a single frequency list, however, they are projected into the same integer, as represented by frequency *1*, *2*, *3*, etc.

The fact, that we are not able to differ finely enough and

to establish the rank numbers properly according to the real measure of usage, does not mean that there exists no usage ordering at all. If many frequency lists would be confronted, or great corpora analyzed, there would arise not only theoretical, but a real chance to distinguish between any two words as to their measure of usage. And, subsequently, it would be possible to establish their different probability rank.

Therefore, the above introduced equation (4) could be also applied (or, it should hold also) at the ending of the frequency list.

*

Similarly, the real representation of x is known and it can be found in every frequency dictionary or frequency list too, for x in the form of the ratio

$$x \approx \frac{\textit{number of occurrences of the word with the measure of usage equal to } x}{\textit{number of all (current words}} = \frac{f_x}{V} \qquad (5)$$

Here again, as the rule, at the beginning of the frequency list x is obviously different for different words, and, at the ending of the frequency list x takes the same value for many words. And again, this empiric knowledge is the typical case of the single frequency dictionary or single frequency list only. More general view, which would deal with many frequency lists e. g. within corpora, would be able to distinguish each corresponding value for each word as finely as requested.

Using the introduced points of view of representations of $N(x)$ and x, we could write

$$r_x\ /\ V = 1 - F\ (\ f_x\ /\ N\ )$$

$$F\ (\ f_x\ /\ N\ ) = 1 - r_x\ /\ V \qquad (6)$$

The question concerning the explicit form of the function $F(x)$, which appears now logically, opens an endless space for sophisticated suggestions, hypotheses, estimations or subjective trials, as an answer.

The following one of them may correspond with thousands of experimental observations, which have been described and expressed with the help of the famous Zipf formula, more directly, than some others.

If we admit for the distribution function (1) the following form of power function

$$F\ (\ x\ )\ =\ E\ .\ x^{-w} \qquad (7)$$

for $x \in \langle 0;1 \rangle$, E, w being constants. Then, after some algebraic steps, having used the above mentioned equation (6), we can get

$$f_x\ =\ const.\ (\ r_x\ +\ R\ )^{-g} \qquad (8)$$

for f_x = frequency of the word with the measure of usage x, where r_x is the rank of this word and *const.*, R and g are constants.

This expression equals to the well known correction of the Zipf formula, as empirically suggested by Mandelbrot (1954) according to other empirical observations. Here it has been deduced by purely stochastic considerations, however, with the help of one presumption (7) only, concerning the distribution of probability of probabilities.

I do admit that it does not sound well to speak about *probability of probabilities*. But, may be, there was not other reason but just the un-elegancy of this expression which caused too small interest of linguists to think of it and to deal with it in order to try one of the ways how to stochastize the Zipf formula.

G.K. Zipf's conception of language as an early prototype of synergetic linguistics

Claudia Prün

pruen@ldv01.uni-trier.de
FB II, LDV/CL, Universität Trier, D-54286 Trier

Topical paper

Abstract

The paper presents a collection of linguistic hypotheses from the work of G. K. Zipf (1902-1950) and reinterprets them as belonging to a systemic conception of language. Frequency of linguistic units as the central concept is linked with the units' size, age, polylexy, semantic specifity and degree of crystallization. The economic constituency principle affects the characteristics of linguistic units and the order parameters of linguistic levels.

0. Introduction

George Kingsley Zipf, born 1902, is most famous for his ranked distributions which everyone might have come across one time or the other. What has entirely fallen into oblivion is his view on language that forms the basis on which these distributions become meaningful. From his publications between 1929 and his untimely death in 1950, I have tried to reconstruct his linguistic hypotheses apart from his curves. Very soon I found that what was assembled there, corresponded to what we now call linguistic synergetics, and that his systemic view was quite consistent.

What I want to present here is a central part of Zipf's linguistic conception, together with a few annotations and the evidence of systemic consistency in this conception.

1. Frequency as the central concept

Zipf's earliest publication bears the title "*Relative Frequency as a Determinant of Phonetic Change*" (Zipf, 1929). I will not touch on his phonological hypotheses here, but relative frequency is central to his work throughout. He compares the language user to an artisan who uses his tools for different tasks. The number of tools, their specialized design for certain jobs, the size of the tools and their distance from the workplace — all depend on or are organized according to the frequency with which any task has to be acheived. Similarly, the inventories of linguistic units are structured according to the requirements of language use. The frequency of linguistic units thus regulates their size, age, semantic specifity and other features of the linguistic system as well as the economy of usage regulates the frequency structure (rank-frequency distribution) of the units on one linguistic level.

2. Size of linguistic units

What Zipf says with respect to sounds is, that the size of the linguistic unit depends strongly on its frequency of use: "***Principle of Frequency***. *The accent, or degree of conspiciousness, of any word, syllable, or sound, is inversely proportionate to the relative frequency of that word, syllable, or sound, among its fellow words, syllables, or sounds, in the stream of spoken language. As usage becomes more frequent, form becomes less accented, or more easily pronounceable, and vice versa*" (Zipf, 1929: 4). For example, it can be shown that strong increase in the use of some sound, e.g. induced by increased use of some affix that sound is weakened and sound shift is triggered (Birkhan, 1979). Also well known is the relation of word frequency and word length (e.g. Köhler, 1986). The relation is inverse: when a unit's frequency grows, its size — however that is to be measured — decreases, and vice versa (Zipf, 1930). We indicate this by a minus sign in the graph. One reason for this relation is economy of language production (minimization of production effort, *minP*). But the less frequent units which carry more information have to be made more conspicious — larger — in order to be transmitted safely. This maximization of conspiciousness (*maxC*) requirement, as we call it following Zipf (1929, 1932 etc.) is obviously antagonistic to *minP*. The simultaneous working of those two "forces" leads to a dynamic equilibrium between frequency and size of a linguistic unit.

3. Age of words

Since the notion of age seems only sensible for words, we will limit ourselves to the lexical subsystem for the moment. Zipf (e.g. 1949: 111) found that among the most frequent words there are also the etymologically oldest words. The ratio of younger words increases with decreasing frequency. We indicate this by a "+" for the balance relation between word frequency and age. Frequent use stabilizes the existence of words and prevents them from disappearing from the lexicon. We will therefore introduce the system requirement *Stab*. On the other hand, the requirement of adaptation *Adap* leads to the introduction of new words and decay of old words, but obviously the effect of adaptation is stronger where words are less frequently used.

The exact mathematical form of the relation has been developed by Arapov (Arapov & Cherc, 1983), much later, but it confirms Zipf's previous hypothesis. We also see that older words tend to be shorter than longer words, the inverse relation indicated by the "-" sign between word size and age. This is consistent with the indirect relation via frequency, because if we "collect" minus and plus signs and multiply them, the resulting sign on any path between two system variables should be the same.

4. Semantic aspects

4.1. Number of word meanings

In Zipf's view, the meaning of a linguistic unit corresponds to the functionality of a tool in the artisan's workshop. The artisan wants to use a small number of different tools as often as possible, to have an economical workplace and not so much changing of tools — i.e. the

unification tendency "all functions on one tool". Without specifying the exact system requirement that leads to the unification tendency here, let us introduce the unification "requirement" *Unif*. Therefore the number of meanings of a word will tend to be larger, the more frequently the word is used (e.g. see Zipf, 1949: 27ff.). On the other hand, the hearer can differentiate best what is meant when every meaning is conveyed by its own signal, resulting in the diversification ("every function its own tool") of words. Again let us cut short here by introducing a "diversification requirement" *Divs* which will of course have to be replaced by some linguistic system requirement notion. The impact of the antagonistic forces of unification and diversification results in a direct equilibrium relation between word frequency and number of meanings, indicated by the "+" sign.

4.2. Specifity

In quantitative linguistics literature we often come across "specifity", quantified in number of meanings. Zipf also sometimes seems to understand it as that. But by careful reading, it appears that we should keep specifity apart from polylexy, as Köhler (1986) coined the expression for a word's number of meanings. Specifity is rather a question of level of the denoted category (Köhler and Altmann, 1993) and we should therefore not investigate specifity of linguistic expressions but of the notions expressed, of the meanings themselves. Zipf was not aware of the necessity to define specifity seperately for the cognitive level, but his understanding of cognition as functional classification of sensory experiences (Zipf, 1935: 287ff. and 1949: 106ff.) implies this conception.

If a linguistic unit is more frequent, it (or rather, its meaning) is not so specific: it can be used in many different contexts, while less frequent units have more specific meanings. Now, what has been said about specifity, frequency and size of linguistic units can also be interpreted for the linguistic levels themselves, or as relations among their order parameters. As we go through the linguistic levels, the most frequent units are the sounds. They are small in size and are rather unspecific as to their meaning (no semantic "meaning", but of course they are functionally classified as to their meaning differentiation function and others). Word, sentences, or texts appear less frequent, are larger and denote much more specific notions. These relations are of course a result of the constituent structure of language.

5. Degree of crystallization

The systagmatic aspect, and the economizing effect of the constituency principle is another great topic in Zipf's work. The more frequently a linguistic unit is used, the more crystallized it becomes. That means, its subunits are more strongly attached to another. We come across this process when morphemes that have once been lexical units end up as affixes, or when strongly crystallized syntactic units are identified as idioms. But again, across the linguistic levels, words are more strongly crystallized than phrases, phrases are — in the mean — more crystallized than sentences of texts. Again the relation emerges in the tension between unification and diversification "forces" (most generally said). Moreover, what is more strongly crystallized tends to be less specific, an inverse relation indicated by the "-" sign.

As with specifity, we can interpret this system element as an order parameter of linguistic level as well as a characteristic of linguistic units.

6. Combination of linguistic units

Another important order parameter is of course the number of units in the inventory of any linguistic level. It stands in close connection to the overall specifity and the order of magnitude of size and freuqency of the units on this level. The constituency principle leads to ever growing inventories, frequencies, sizes, specifity etc. when we move through the levels from low to high. (Of course every system element or characteristic has to be defined seperately for every level, if it is to be measured, but such was Zipf's idea of language.) We introduce the inventory size here, though Zipf (1949: 67ff.) only has the principle of "inventory minimization".

Zipf (loc.cit) again considers frequency to be the driving power behind the processes of combination. He supposes that the more frequent a lower level unit is, the more different higher level units it will be part of, which in turn will tend to be larger, be used more frequently and are probably more strongly crystallized. The low specifity of the unit is prerequisite or even the cause for it being used so often in more specific higher level expressions.

The assumption of stronger crystallization with more frequent constituents can easily be shown on the example of German verbal prefixes. While the most frequent prefixes are inseparable, only less frequent prefixes can usually be separated from the verbal stem (Zipf, 1935: 148).

What is also clearly correct, is the growing construct size with the growing constituent frequency. Today, we attach growing construct size to decreasing constituent size, but obviously on all levels unit sizes grow with decreasing frequency, resulting in the relation that Zipf assumed. It is well established under the name of "Menzerath's Law" (for a bibliography, see Prün, 1994).

In assuming that construct frequency will tend to grow with increasing constituent frequency, Zipf seems to be mistaken, though. This is certainly not true for levels of immediate constituency, as we can see simply from our graph, which becomes inconsistent. Any path between unit frequency and unit size on one level should yield "-", but this path results in a "+". In addition, it is problematic because of course units of high freuqency are usually combined with rarer units, and the constituents' frequencies probably don't much influence the construct's frequency. (We too do not know if we are to measure the constituent frequency in a text or e.g. its occurence in the higher level inventory.)

The first mentioned implication, participation in a greater number of different constructs with increasing frequency is accounted to the units' small specifity. We suspect, though, that here the mathematical form needs not be monotone, as with the other relations, depending on measuring and on functional aspects of the levels and units concerned. This is still open to statistical testing.

7. Conclusion

We have now limited ourselves to only a part of Zipf's linguistic hypotheses. We can see

that except for one point, his systemic view of language was quite consistent. It still bears many points that challenge us to work on linguistic theory, especially the relations of linguistic levels, and on the quantitative rendering and measurement of linguistic entities.

Bibliography (selected)

Arapov, Michail V. und Cherc, Maja M. (1983): *Mathematische Methoden in der historischen Linguistik*. Quantitative linguistics 17. Bochum: Brockmeyer.

Birkhan, Helmut (1979): *Das „Zipfsche Gesetz", das schwache Präteritum und die germanische Lautverschiebung*. Sitzungsberichte der österreichischen Akademie der Wissenschaften, philosophisch-historische Klasse 348.

Köhler, Reinhard (1986): *Struktur und Dynamik der Lexik*. Quantitative linguistics 31. Bochum: Brockmeyer.

Köhler, Reinhard und Altmann, Gabriel (1993): Begriffsdynamik und Lexikonstruktur. In: Beckmann, Frank und Heyer, Gerhard (Hrsg.): *Theorie und Praxis des Lexikons*. Berlin, New York: DeGruyter, 173-190.

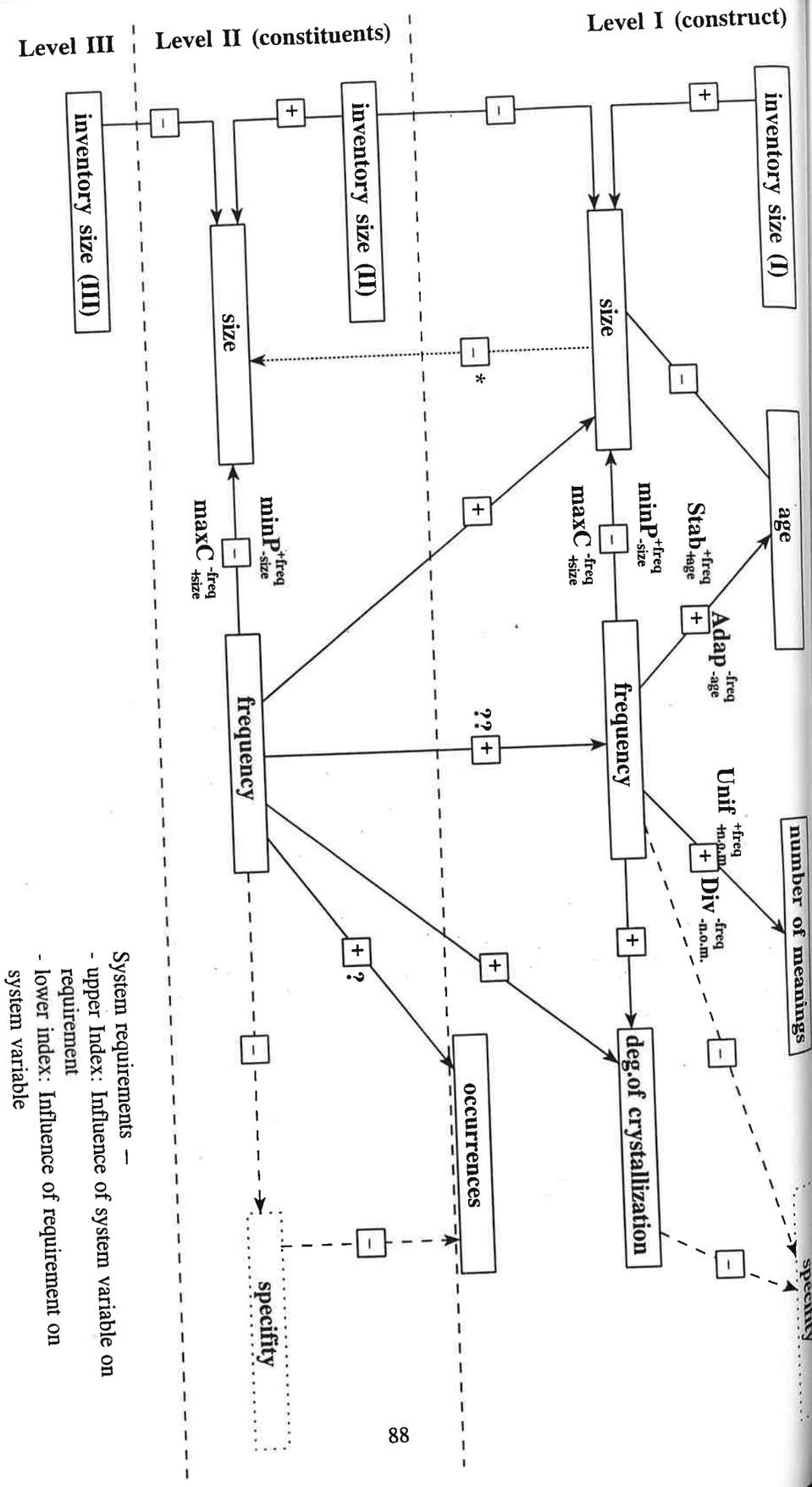
Prün, Claudia (1994): Validity of Menzerath-Altmann's law: Graphic representation of language, information processing systems and synergetic linguistics. In: *Journal of quantitative linguistics* 1, 148-155.

Zipf, George Kingsley (1929): *Relative frequency as a determinant of phonetic change*. Harvard studies in classical philology 40.

Zipf, George Kingsley (1932): *Selected studies of the principle of relative frequency in language*. Cambridge/Mass., Harvard Univ.Press.

Zipf, George Kingsley (1935): *The psycho-biology of language. An introduction to dynamic philology*. Cambridge/Mass., M.I.T. Press, 2nd ed. 1968 [1st. edition: Boston, Houghton-Mifflin, 1935].

Zipf, George Kingsley (1949a): *Human behavior and the principle of least effort. An introduction to human ecology*. New York: Hafner reprint, 1972. [1st. edition: Cambridge/Mass., Addison-Wesley, 1949.]



preliminary version of paper for Qualico-97

title: **Natural statistics in language modeling**

author: **Royal Skousen**

Department of English
Brigham Young University
Provo, Utah 84602 USA

royal_skousen@byu.edu

short summary:

Natural statistics permit us to predict linguistic behavior statistically, but avoid any direct knowledge of probability distributions. Such statistics have the ability to predict stochastic language behavior as if the underlying probability distribution is known.

topical paper, long (20 minutes plus 10)

topic area: "methodological problems of linguistic measurement, model construction, sampling and test theory" (closest one)

The crucial problem in analogical descriptions of language is to locate heterogeneity in the contextual space. One of the major innovations in Skousen 1989 and 1992 is the notion of a natural statistic. Traditional statistical tests require knowledge of either the underlying probability distribution for the test or a distribution that approximates the underlying distribution. Unfortunately, such tests are mathematically very complex and completely unsuitable as psychologically plausible models of decision making. A natural statistic, on the other hand, avoids any direct consideration of probability distributions, yet has the ability to predict stochastic behavior as if the underlying probability distribution is known.

Two natural statistics have been discovered thus far. The first natural statistic is based on the rate of agreement, which derives from a quadratic (not a logarithmic) measure of uncertainty. The decision rule for determining heterogeneity is a very simple one: *maximize the rate of agreement.* This decision rule is a very powerful one, with a level of significance near one-half. Smaller levels of significance (at 0.05 or less) can also be defined in terms of this statistic, so the test can be made fully equivalent to standard statistical tests.

The second natural statistic is, on the surface, an incredible one in that it eliminates the need for any mathematical calculation at all. By simple inspection, all cases of *potential heterogeneity* in the contextual space are eliminated. This test represents the most powerful test possible: any context that could possibly be heterogeneous is declared to be heterogeneous. The decision rule for this statistic is extremely simple: *minimize the number of disagreements.* Such a powerful test is, of course, completely contrary to all standard statistical procedure, but by adding the concept of imperfect memory, this natural statistic gives the same basic results as standard statistics. In fact, there is a direct correlation between imperfect memory and level of significance: the more imperfect the memory, the smaller the level of significance. We always use the most powerful test based on minimizing the number of disagreements, but test at a more typical (that is, smaller) level of significance by randomly selecting only a small fraction of the data. In other words, a "statistically significant" relationship is one that holds even when most of the data is forgotten.

In this paper I will develop this second (most powerful) natural statistic and apply it to several statistical problems that arise in predicting language behavior, including the following:

(1) estimating an unknown probability by assuming that the probability of remembering a particular occurrence equals one half;

(2) determining the most frequent outcome, again by assuming the same level of imperfect memory (that is, one half);

(3) comparing standard discrete multivariate analysis (as in Bishop, Fienberg, and Holland 1975) with the indirect approach of analogical modeling.

Speakers have the ability to estimate frequencies of occurrence, predict which outcome is the most frequent, and use language as if speakers had determined the statistical relationships between various linguistic variants. Within a psychologically plausible theory of analogical modeling, natural statistics allow speakers to make such judgments without requiring them to posit highly complex statistical distributions or to directly calculate probabilities mathematically.

References

Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice* (Cambridge, Massachusetts: MIT Press).

Skousen, Royal (1989). *Analogical Modeling of Language* (Dordrecht, The Netherlands: Kluwer Academic Press).

Skousen, Royal (1992). *Analogy and Structure* (Dordrecht, The Netherlands: Kluwer Academic Press).

Syntactic structures: properties and interrelations

(QUALICO 97)

Reinhard Köhler, Trier

ABSTRACT

There have been few investigations of quantitative properties and dependences of syntactic units so far, despite the relevance of the syntactic level of linguistic analysis. Existing theoretical results, viz. Menzerath-Altmann's law (applied to the sentence level), and empirical findings such as sentence and clause length distributions, have not yet been integrated into a common explanative model.

In the present paper, a first step is taken towards setting up a set of hypotheses which is compatible, with respect to its theoretical foundation, with the approach of systems theoretical (synergetic) linguistics. In particular, system requirements of language economy are taken into account. By means of empirical data from an English text corpus it is shown that syntactic constructions and elements follow patterns similar to those of the lexicon and morphology.

Finally, some consequences of these findings for psycholinguistic research and practical applications for parsing strategies are discussed.

GRAMMAR

Third International Conference on Quantitative Linguistics, Helsinki

Submission for a Presentation
by Marc Hug, Strasbourg
Address : 19, rue Oberlin, 67000 Strasbourg (France)
Tel. (33)3 88 35 08 55
e-mail : hug@ushs.u-strasbg.fr

Topic : Syntax and Semantics, related to statistical features of textual corpora

Title : The French demonstrative particles *-ci* and *-là* : linguistic intuitions and statistical facts.

The French system of Demonstratives includes four couples of units that finish with *-ci* and *-là* respectively, which are the following :

- the verbal forms *voici* and *voilà* (traditionally, dictionaries and grammars call them either "prepositions" or "adverbs" or "presentatives" or "introducers", but their distributional characteristics show that they are verbs) ;
- the discontinuous noun determiners *ce* N-*ci* and *ce* N-*là* ; these are different from the three other couples by the fact that the demonstrative determiner *ce* may be used alone, without *-ci* or *-là*, while the others are not usable, or do not have a demonstrative meaning if used without the particles ;
- the masculine / feminine pronouns *celui-ci* and *celui-là* ;
- the neutral pronouns *ceci* and *cela*.

Almost all French grammars, from the most traditional grammars to the most "modern" linguistic ones, present approximately the same analysis of the two particles *-ci* and *-là* :

- *ci* defines a proximal demonstrative value, *là* a distant demonstrative value ; the "distance" may be either in space or in time, or in the number of words between the demonstrative and its antecedent ;
- where *ci* and *là* do not imply any notion of distance, *ci* refers to a following term, *là* to a preceding one.

Grammars also agree in emphasizing the correlated use of both types, i.e. *ce* N-*ci* as opposed to *ce* N-*là*, *celui-ci* as opposed to *celui-là*, *ceci* as opposed to *cela*, *voici* as opposed to *voilà*. Finally they all agree in stating that *là*-forms, at least in everyday-speech, are increasingly preferred at the expense of *ci*-forms.

The data provided by three large textual corpora (shortly : (1) Maupassant, (2) Proust, (3) various texts of the period after 1960) do not confirm these statements :

- The *-ci* forms have not the same relative frequency in all four couples of forms ; the determiner and the neutral pronoun have enormous amounts of *-là* forms and only few occurrences of *-ci* forms, while the other two couples show more balanced occurrence numbers ;
- They do not show parallel evolutions in all couples between 1880 and 1990 ; while *voici* seems to become less usual in comparison with *voilà*, the *-ci* forms take increasing percentages in the other three form couples ;

- This means that, globally, *-ci* forms seem to be much more frequent in the recent corpus than in the oldest one, what could be in contradiction with the widespread opinion that speakers more and more avoid *-ci* forms in everyday speech.

But these observations raise several problems, as the following :

- If there is a coherent semantic analysis that can apply to each of the particles, what can explain the fact that the relative frequencies of both are far different from one couple to the others ? Does the semantic content of the rest of each unit provide the clue for an explanation ? For example, could the normal functioning of anaphora provide an explanation for the fact that the neutral pronoun (which mostly refers to a propositional content) is much more used in the *cela* form than in the *ceci* form, while the masculine-feminine pronoun, which normally anaphorizes a noun phrase, is now more used in written French under the *-ci* form *celui-ci, celle-ci* etc. ?
- The linguistic intuitions of the grammarians may be considered as belonging to the conscious part of their linguistic competence ; how does it come about that they think the *-ci* particle to become less frequent when its frequence, on the contrary, seems to be increasing ? To what extent must we explain the observed percentages of *-ci* forms by the special stylistic or semantic character of the different texts ? Is the linguists' intuition related to the fact that the common oral speech prefers *-là* forms much more strikingly than does the written language, and with the fact that written language is perceived as a more traditional or archaic form of the language than oral speech ?
- What kind of factors could be invoked if we want to explain the various changes that are observed concerning the frequencies of these demonstratives ? E.g., the fact that *voici* and *voilà* are both much more frequent in the Maupassant corpus than in the two others will be easily explained by the narrative character of the first and the more descriptive character of the two other corpora. It is not sure that all the differences are as easy to explain.

The contribution will present these problems more precisely and expose a tentative solution for each of them. The thorough examination of the three corpora is still in progress.

Christiane Hoffmann
Trier University
Rechenzentrum
Universitätsring 15
D-54286 Trier
email:hoffmanc@uni-trier.de

Word order and the principle of "Early Immediate Constituents"

Abstract: Embedded in functionally orientated linguistic modelling and typologic argumentation EIC seeks to explain word order variation on phrase level in performance and processes of grammaticalization on the basis of human information processing. By quantification and empirical testing of EIC driven hypotheses the principle proves rather plausible although some questions are still to be answered and integration into a wider linguistic model will have to be accomplished.

The principle of Early Immediate Constituents (EIC) states that the ordering of phrases in an utterance/sentence is governed by the length of the phrases (Hawkins 1992, 1994) and constitutes a scientific elaboration of an idea that has been around for quite a while.

Embedded in functionally orientated linguistic modelling and typologic argumentation EIC seeks to explain word order variation in performance and processes of grammaticalization on the basis of human information processing.

In previous work by Hawkins (1994), Siewierska (1991) and Hoffmann (1995) it could be shown by quantification and empirical testing of EIC driven hypotheses that the principle proved rather plausible in various languages with regard to different syntactic constructions.

First steps to quantify the principle in such a way that it is apt to empirical testing and to integrate the principle in synergetic models (similar to the one elaborated in Köhler, 1986) have been taken.

These results incite new questions to study, e.g.:

1. Aspects concerning the mathematical model of the distributions are to be explored: E.g. in modelling the preference for extraposition of heavy subjects the length of the verbal phrase could be accounted for.

2. With regard to the ordering of two adjacent prepositional phrases a weighted quantification should be tested on a larger sample.

3. Which operationalization of "heaviness" should be chosen?

4. As to the question of the linguistic reality of the models' parameter(s) genre-differentiated texts should be studied in large samples.

5. Is there a way to couple EIC-driven hypotheses about intra-phrase ordering (Uhlířová, to appear) with EIC-driven hypotheses about inter-phrase ordering?

Furthermore an essential question has to be raised which concerns the principle plausibility of EIC, i.e. a blind spot in the logical stringency of the functioning of the principle. EIC implies that the hearer "knows" with the first word of the last phrase of the construction in question that there are no more phrases to follow. This knowledge allows her - according to the underlying mechanism that is presupposed - to free working memory from phrase level information early leading in turn to relieved information processing. But how can she be sure that indeed no more phrases follow? This blind spot calls for a modification/enhancement of

the supposed underlying information processing mechanism. The need for a supplementary constraint seems obvious:

Working memory can only be freed on the rather sound assumption that the probability for another phrase to follow is supposed to be low or near zero (for the idea cf. e.g. Köhler, 1984). This implies previous tuning in form of syntactic input during language learning and exposure. For empirical testing this tuning might be modeled by a probabilistic grammar.

References

Hawkins, John A.: A performance theory of order and constituency. Cambridge, Univ. Pr. 1994 (Cambridge studies in linguistics ; 73)

Hawkins, John A.: Syntactic weight versus information structure in word order variation. In Jacobs, Joachim (Ed.): Informationsstruktur und Grammatik. Opladen: Westdt. Verl., 1992 (Linguistische Berichte / Sonderheft ; 4)

Hoffmann, Christiane: Quantitativ-funktionalanalytische Untersuchungen zur Wortstellungsvariation. Magisterarbeit, Trier, 1995 (Unpublished M.A. thesis)

Köhler, Reinhard: Zur Interpretation des Menzerathschen Gesetzes. In: Boy, Joachim (Ed.): Glottometrika 6. Bochum: Brockmeyer, 1984 (Quantitative linguistics ; 25)

Köhler, Reinhard: Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. Bochum: Brockmeyer, 1986 (Quantitative linguistics ; 31)

Siewierska, Anna: Syntactic weight vs information structure and word order variation in Polish. IN. Journal of linguistics 29(1993), 233-265

Uhlířová, Ludmila: Length vs order: On the word length and clause length form the word order perspective. To appear.

Measuring Linguistic Complexity : The Morphological Tier

Patrick Juola*

A standard question in freshman linguistics is "Which language is the most complex," with related questions about the various aspects of complexity, typically the ones the student is having trouble learning. Aside from its armchair interest, it can also be an important typological question, as well as possibly shedding light on the human brain's processing of linguistic information. Unfortunately, there is no accepted method for measuring and comparing such aspects of complexity, and linguists are reduced to answers based more on politics than on empirical evidence. Morphological complexity, in particular, is an obvious testbed for any theories about the possibility of getting these measurements; it is intuitively obvious that some languages (for example, Finnish) are "morphologically complex" while others are more simple. On the other hand, claims about (e.g.) semantic differences are less intuitive and less widely accepted.

One of the most comprehensive works on morphological complexity is that of Nichols[4]. In her work, she develops a measure of complexity based on the number of points at which a typical sentence is capable of receiving inflection. This is one of a very few, and perhaps the only, attempt to produce a comparative numerical index of complexity. She further applies this analysis to a sample of nearly 200 languages, some findings of which are replicated below. It is significant that she does not attempt to justify this index in mathematical terms, since there is no well-accepted standard against which to compare this. It is into this near-vacuum, then, that one tries to develop a theory to support numerical validation of this sort of measurement.

Juola[3] proposes a functionalist approach to this sort of measurement, based on the information-theoretic concept of "information contained in a sample of text." The mathematics of information theory (for more details, consult [1, 7]) can be summarized in the basic idea that "information" equates to the unexpectedness of a piece of information and the degree to which something cannot be simply predicted from other aspects of the sample. By examining (translations of) the same text for their information content (as measured by standard compression techniques, e.g. [8]), one can arrive at an overall measure of "linguistic complexity" for an entire language, rather than for the small structures measured by [2, 5].

We attempt here to extend this analysis and to determine whether or not smaller-scale factors than "a language" can be thus analyzed. Restricting attention for the moment to the morphological tier, we attempt to isolate one factor, the complexity contained at the so-called "morphological tier." A morphologically complex language, under this view, is simply one where the information conveyed by morphological processes contributes substantively to the information conveyed by the entire text: for example, one where the agent/patient relationships cannot be determined by examination of the word order.

This complexity can be investigated by a careful manipulation of the morphological information in a sample. A simple approach to preparing such "morphologically degraded" texts is to replace every word type with an arbitrarily chosen (random) symbol; this has the effect of replacing the regularities at the morphological tier with random (unpredictable, and hence maximally "informative") noise. This rewriting process will have two main effects. First, information at the phonological tier is irrevocably destroyed; this results in a net loss of information and corresponding overall sizes of compressed files. Second, relationships between and among words at the syntactic tier (and above) are unchanged; the primary effect is to make the prediction of particular word-forms *on the basis of other word-forms in the sentence* more difficult; i.e. to inflate the information content at the morphological tier. By comparing ratios of the information contained in the raw samples with the information contained in the morphologically degraded samples (measured as in [3] by size of compressed text samples), one can achieve an

*Oxford University, Oxford, UK, patrick.juola@psy.ox.ac.uk

Table 1: Size and information content (in bytes) of various samples

Language	Uncompressed	Comp.(raw)	Comp. ("cooked")	R/C Ratio
Dutch	4,509,963	1,383,938	1,391,046	0.994
English	4,347,401	1,303,032	1,341,049	0.972
Finnish	4,196,930	1,370,821	1,218,222	1.12
French	4,279,259	1,348,129	1,332,518	1.01
Maori	4,607,440	1,240,406	1,385,446	0.895
Russian	3,542,756	1,285,503	1,229,459	1.04

approximate numerical measurement of morphological complexity.

This experiment has been performed using similar Biblical texts to Juola's. [6] provides machine-readable copies of the Bible in a variety of languages and translations, including Dutch, English, Finnish, French, Maori, and Russian. The sample taken from each language was the entire text of the Bible (Old and New Testaments, but excluding Apocrypha), approximately 825,000 words in English or about 4.3 megabytes of raw text.

The results are attached as table 1. As can be seen, the resulting r/c ratios sort the languages into the order (of increasing complexity) Maori, English, Dutch, French, Russian, Finnish. It is also evident that there is significant (phonological) information which is destroyed in the morphological degradation process, as three of the six samples actually have their information content *reduced*.

The findings above are largely unsurprising; few would quibble with the statement that some languages are more morphologically complex than others, or that Finnish and Russian are complex compared to English. However, they provide further evidence in support of the usefulness of Juola's general information-theoretic approach. They also demonstrate an empirical and objective approach to the direct measurement of something (morphological complexity) previously subjective.

This method of selective alteration (or deletion) of tiers can presumably be extended to other areas; for example, by selectively altering the syntactic tier, one could produce similar measures for the syntactic complexity of a given language. Much further work is clearly required, both to refine the tools used and to more carefully measure their accuracy, as well as to determine the areas and extent of their usefulness. This work only touches the surface of what may prove to be a very wide-ranging and fruitful area of study.

References

[1] Norman Abramson. *Information Theory and Coding.* McGraw-Hill, New York, 1963.

[2] Brent Berlin and Paul Kay. *Basic Color Terms : Their Universality and Evolution.* University of California Press, Berkeley, CA, 1969.

[3] Patrick Juola. Measuring linguistic complexity. Under review, 1996.

[4] Johanna Nichols. *Linguistic Diversity in Space and Time.* University of Chicago Press, Chicago, IL, 1992.

[5] Revere D. Perkins. *Deixis, Grammar and Culture*, volume 24 of *Typological Studies in Language.* John Benjamins, Amsterdam, 1992.

[6] Larry Pierce. *The ONLINE BIBLE User's Guide.* Woodside Bible Fellowship, Ontario, Canada, 1993.

[7] Claude Elmwood Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(4):379–423, 1948.

[8] Jakob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, IT-23(3):373–343, May 1977.

MULTIVARIATE RULE ANALYSIS OF CASE ASSIGNMENT IN BILINGUAL DISCOURSE.

Svitlana Budzhak-Jones

University of Ottawa

Mailing address: 1 East Water Street, Lock Haven, PA, 17745 USA.
E-mail: ljones@eagle.lhup.edu.

Summary. The assignment of case to nouns of different status in Ukrainian-English bilingual discourse is analyzed by variable rule analysis, Goldvarb 2.0. It is demonstrated that the factors which condition structural case are identical, irrespective of the languages involved. However, inherent case assignment patterns differently, owing to the peculiarities of the specific language system which produced the noun.

Topical paper in sociolinguistics and syntax.

Grammatical relations between nouns in any language are expressed by the category of case (Blake 1994). Within Government and Binding theory, it has been argued that some of these grammatical dependencies are structurally determined by a Universal grammar, whereas others are language specific, i.e. inherent (Chomsky & Lasnik 1991). Structural case is assigned to a noun phrase (NP) according to its position in a structural configuration under government and can be overtly or covertly realized. Inherent case is peculiar to a particular language and has to be specified in the lexicon.

In the following paper we will examine the mechanisms of case assignment to nouns within a bilingual discourse involving languages with distinct case systems. Based on the assumption that loanwords are fully syntactically, morphologically and (sometimes) phonologically assimilated into the host language (Poplack 1993) we will expect that they will obey the rules of case assignment in exactly the same way as their native counterparts, whereas nouns which are code-switched will retain their original grammar and will not submit to the same rules of case assignment in the same manner as host language nouns. We will base our research on two distinct languages, Ukrainian and English, used simultaneously in bilingual discourse.

English has a three-case system (Quirk et al. 1980). Nominative and accusative are assigned structurally and remain unmarked[1], whereas genitive (or possessive), is morphologically marked. Verbs and prepositions are accusative case assigners in English. The abstract element INFL assigns nominative to subjects, and genitive is assigned by nouns inherently (Chomsky 1986:194). Quantifiers, numerals and adverbs neither receive a

[1]With the exception of some pronouns, which are overtly marked (e.g. me, him, them).

case nor assign it themselves (cf. Haegeman 1992:162). Moreover, there is no case agreement in English.

Ukrainian[2] has seven cases (Pljushch 1994). Similar to English, structural determined cases, nominative and accusative, are assigned in exactly the same way. However, unlike English, in Ukrainian accusative case may be overtly realized (feminine and animate masculine nouns). All other cases are language specific and most of the time have an overt case morphology (with the exception of some plural nouns in genitive). These morphological cases can be assigned by verbs, prepositions, nouns, quantifiers, some adverbs, as well as an empty category (in case of adjuncts). Finally, unlike English, every NP head in Ukrainian must establish a case concord with its modifiers.

Our research is based on the data collected by the author in the Ukrainian-English bilingual community in Lehighton, Pennsylvania (USA), and comprises 36 hours of natural tape-recorded sociolinguistic interviews with 25 bilingual speakers. For this project two corpora are employed:[3] 1) monolingual Ukrainian (1951 tokens) and English-origin (1637 tokens) nouns, used in the otherwise Ukrainian context. Both were extracted from the same interviews of the same informants (see Budzhak-Jones (1996) and Budzhak-Jones (in preparation) for extensive discussion of Lehighton data base).

All nouns were coded for a number of factors which could have influenced case assignment. These include: 1) syntactic position, i.e. structural or inherent; 2) case assigner by feature, i.e. the ability to assign a syntactic and/or morphological case; 3) case assigner by type, i.e. verbs, prepositions and other; and 4) case agreement. These factors were then tested on their significance in influencing a non-standard case marking, with respect to prescriptive rules of literary Ukrainian, as inferred from Ukrainian grammars (e.g. *Ukrajins'kyj pravopys* (Ditel' 1993), *Suchasna ukrajins'ka literaturna mova* (Pljushch et al. 1994), etc.).

The data was analysed by variable rule analysis, Goldvarb 2.0 for Macintosh (Rand and Sankoff 1990). This is a multiple regression procedure which extracts regularities from naturally occuring frequencies in the corpus-based data. It makes an assessment of the influence of different factors on a particular choice, and retains the most statistically significant factors which increase the likelihood of a dependent variant to occur. It is performed on two levels (Sankoff 1988). The step-up procedure tries to find a single statistically significant factor-group, and then gradually adds other factor groups to measure their significance. The step-down solution is based on the reversed procedure, where the

[2]Ukrainian is an Eastern Slavic language with fusional morphology.
[3]Monolingual English corpus was not included in this analysis, since there were only two possessive case tokens (see Budzhak-Jones, in preparation). All other cases were structural and null marked.

likelihood of the occurrence of the dependent variable is calculated first and then factor-groups are eliminated one-by-one starting from least significant. Finally, both steps retain the most significant factors influencing a given choice. If the factors considered in the analysis are not entirely independent, a one level calculation can be executed, which analyzes the input of all groups simultaneously.

Based on the case assignment features of both interacting languages, we shall therefore anticipate that the case assignment to lone English-origin nouns used in the otherwise Ukrainian discourse, will be influenced by the same factors and to the same extent as their monolingual Ukrainian counterparts, if the former are borrowed. English-origin nouns which retain their English grammar, and hence are code-switched, will not parallel the behavior of native nouns, and will be conditioned by different factors with respect to case assignment.

The results of our variable rule analysis are shown in Table 1. Non-standard case marking in monolingual Ukrainian nouns highly depends on three factor-groups: case position, a case assigner's feature, and its type. English-origin nouns are influenced by the first two factors, but not the last one. Moreover, the hierarchy of effect is the same across the corpora. Nouns in both corpora are most likely to receive a non-standard case marking when the inherent case is required, and are less so if the structural case is assigned. When the case assigner has the property to assign both syntactic and morphological cases, the probability of non-standard marking is the highest, whereas when the case is assigned

Table 1. Variable rule analysis of the contribution of factors selected as significant to non-standard case marking in Ukrainian context across corpora.

	Ukrainian monolingual		**English-origin** in Ukrainian	
CORRECTED MEAN:	**.055**		**.107**	
TOTAL N:	**1951**		**1637**	
	Probability	N	Probability	N
Case position				
Structural	**.310**	(1049)	**.337**	(1141)
Inherent	**.717**	(902)	**.825**	(496)
Case assigner (by feature)				
Syntactic	**.262**	(467)	**.400**	(453)
Morphological	**.432**	(503)	**.481**	(245)
Syntactic and morphological	**.653**	(981)	**.554**	(939)
Case assigner (by type)				
Other	**.339**	(197)		
Preposition	**.408**	(675)		
Verb	**.619**	(1079)		
FACTORS NOT SELECTED				
Case agreement	X		X	
Case assigner (by type)			X	

syntactically only, non-standard marking is the lowest. This is not surprising since both syntactic and structural case allow most null morphology in both languages.

Cases agreement did not have a statistically significant effect in both corpora. The type of a case assigner, however, was selected significant only for monolingual Ukrainian nouns. Since it has been demonstrated that single word English-origin tokens used in Ukrainian discourse may be the product of either code-switch or borrowing, depending on their overt morphology (Budzhak-Jones, in preparation), we will next examine the influence of different factors on nouns with overt Ukrainian morphology as opposed to the nouns with a null inflection. Unfortunately, the interaction between different factor-groups when the data was split in two subcorpora (overt and null-marked), prevented us from performing a 2-level analysis. Therefore, we had to execute only a 1-level procedure.

Table 2 shows that all factor groups which were selected significant for the entire corpora (i.e. Table 1), show the same results in both corpora. Moreover, the hierarchies of effect are parallel across both languages. They differ only with respect to case agreement, but this factor was not statistically significant for either corpus in the previous analysis. Moreover, the relative weight between the factors in this factor-group is very small.

Table 2. Variable rule analysis of the contribution of factors selected as significant to non-standard case marking of overtly inflected nouns across corpora.

	Ukrainian monolingual			**English-origin** in Ukrainian		
CORRECTED MEAN:	**.059**			**.088**		
TOTAL N:	**1640**			**803**		
	Weight	Input & weight	N	Weight	Input & weight	N
Case position						
Structural	.349	**.03**	(787)	.431	**.07**	(455)
Inherent	.640	**.10**	(853)	.590	**.12**	(348)
Case assigner (by feature)						
Syntactic	.224	**.02**	(337)	.314	**.04**	(150)
Morphological	.449	**.05**	(472)	.467	**.08**	(176)
Syntactic and morphological	.650	**.10**	(831)	.573	**.12**	(477)
Case assigner (by type)						
Verb	.593	**.08**	(870)	.560	**.11**	(478)
Preposition	.397	**.04**	(608)	.431	**.07**	(285)
Other	.391	**.04**	(162)	.287	**.04**	(40)
Case agreement						
Agreement required	.555	**.07**	(604)	.455	**.10**	(287)
Agreement free	.468	**.05**	(1036)	.525	**.07**	(516)

The influence of different factor-groups on case assignment of null marked nouns is shown in Table 3. The results are quite different. The two corpora are influenced by the same factors in the same manner, only with respect to case position. In all other factor-groups not only is there a considerable difference in the probability of occurrence of non-

standard marking, but also the hierarchy is different within each factor group. This is important evidence that null marked English-origin nouns are not assigned case in the same way as their native counterparts, and therefore, they may be code-switched.[4]

Table 3. Variable rule analysis of the contribution of factors selected as significant to non-standard case marking of nouns with null morphology across corpora.

	Ukrainian monolingual			**English-origin** in Ukrainian		
CORRECTED MEAN:	**.016**			**.159**		
TOTAL N:	**311**			**834**		
	Weight	Input & weight	N	Weight	Input & weight	N
Case position						
Structural	.239	**.00**	(262)	.195	**.04**	(686)
Inherent	.998	**.89**	(49)	.999	**.99**	(148)
Case assigner (by feature)						
Syntactic	.380	**.01**	(130)	.513	**.17**	(303)
Morphological	.138	**.00**	(31)	.424	**.12**	(69)
Syntactic and morphological	.690	**.03**	(150)	.503	**.16**	(462)
Case assigner (by type)						
Verb	.727	**.04**	(209)	.461	**.14**	(651)
Preposition	.099	**.00**	(67)	.567	**.20**	(165)
Other	.165	**.00**	(37)	.962	**.83**	(18)
Case agreement						
Agreement required	.468	**.01**	(136)	.509	**.16**	(284)
Agreement free	.525	**.02**	(175)	.496	**.16**	(550)

Note, however, that nouns in both corpora are almost categorically non-standard in the position where an inherent case is required. We, therefore, performed one more analysis excluding the case position from consideration. This eliminated the interaction within each subcorpus, and allowed us to execute a binomial procedure. The results are shown in Table 4. Nouns with overt Ukrainian morphology irrespective of the word's origin are assigned case within the Ukrainian context in exactly the same manner, and it is significantly conditioned only by the ability of a case assigner to assign case. Nouns with null Ukrainian morphology are conditioned by two factor-groups. With respect to case assigner's feature they show exactly the same pattern, whereas with respect to the type of a case assigner they differ considerably. This may be taken as evidence that our null marked English-origin data are not monolithic by nature. Some nouns in this corpus may be borrowed, whereas others may be code-switched.

To conclude, we have demonstrated that in bilingual discourse nouns show the properties of the grammar by which they were generated. Both native and borrowed nouns are conditioned by the same factors and in the same manner. Code-switched nouns differ

[4] Since we do not have evidence of the behavior of monolingual English nouns with respect to case assignment, we cannot conclude that our English-origin nouns are definitely the product of code-switching.

Table 4. Variable rule analysis of the contribution of factors selected as significant to non-standard case marking of nouns across corpora, excluding the case position.

	Ukrainian monolingual		**English-origin** in Ukrainian	
	Overt	Null	Overt	Null
CORRECTED MEAN:	**.064**	**.043**	**.091**	**.169**
Case assigner (by feature)				
Syntactic	**.181**	**.131**	**.293**	**216**
Morphological	**.508**	**.962**	**.498**	**.994**
Syntactic and morphological	**.644**	**.725**	**.569**	**.523**
Case assigner (by type)				
Verb		**.697**		**.423**
Preposition		**.142**		**.738**
Other		**.177**		**.844**
FACTORS NOT SELECTED				
Case agreement	**X**	**X**	**X**	**X**
Case assigner (by type)	**X**		**X**	

from them and show the statistically different patterns of case marking. Furthermore, our results proved that the properties of a Universal grammar are realized in the same manner across different languages, whereas language specific features are pertinent only to the grammatical system involved in case assignment.

REFERENCES:

Blake, B. J. 1994. *Case*. Cambridge, UK: Cambridge University Press.

Budzhak-Jones, S. 1996. Social conditions and code-switching patterns. In A. Pasquini, et al. (Eds.), *Proceedings of the 1996 Annual Conference of the Canadian Linguistic Association*. Calgary, Alberta: Calgary Working Papers in Linguistics. 23-34.

Budzhak-Jones, S. V. in preparation. *Little things mean a lot: Single-word incorporations in Ukrainian- English bilingual discourse*. Ph.D. dissertation. University of Ottawa.

Ditel', O. A. (Ed). 1990. *Ukrajins'kyj pravopys*. Kyjiv: Naukova dumka.

Chomsky, N. 1986. *Knowledge of language, its nature, origin, and use*. New York: Praeger.

Chomsky, N. & H. Lasnik. 1991. Principles and parameters theory. In J. Jacobs et al. (Eds.), *Syntax: An International Handbook of Contemporary Research*.Berlin: Walter de Gruyter.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1972. *A grammar of contemporary English*. New York: Harcourt Brace Javanovich, Inc.

Rand, D. & D. Sankoff. 1990. GoldVarb. *A variable rule application for the Macintosh*. Version 2. Montreal, Canada: Centre de recherches mathématiques, Université de Montréal.

Pljushch, M. J. (Ed.). 1994. *Suchasna ukrajins'ka literaturna mova*. Kyjiv: Vyshcha shkola.

Poplack, S. 1993. Variation theory and language contact. In D. Preston (Ed.), *American Dialect Research: An Anthology Celebrating the 100th Anniversary of the American Dialect Society*. Amsterdam/Philadelphia: John Benjamins. 251-286.

Sankoff, D. 1998. Variable rules. In U. Ammon, et al. (Eds.), *Sociolinguistics: An International Handbook of the Science of Language and Society*. Berlin: Walter de Gruyter. 140-161.

TEXT ANALYSIS

Dependent observations in text linguistics: the occurrence of the passive in Dutch

Louise Cornelis and Huub van den Bergh
UIL - Utrecht University
Trans 10
NL-3512 JK Utrecht
International Quantitative Linguistics Conference
August 1997 - Preliminary Paper (topical)
louise.cornelis@let.ruu.nl
huub.vandenbergh@let.ruu.nl

0. Summary topic area: dependency/frequency analysis in text linguistics, multilevel analysis.

We will show that a multilevel model of frequency data for the occurrence of the passive in Dutch texts sheds a new light on that occurrence across text genres. Traditional statistical tests show a difference in frequency between genres, which actually has to be attributed to a difference within genres, between texts. We will present the analysis, and discuss its implications.

1. Introduction and background

In text linguistics (and related branches such as functional linguistics), phenomena are often considered to occur independently in discourse. At least, that is the conclusion we can draw when we observe the kind of statistical methods used, and the conclusions drawn from them. For example, Givón (1993:8) reports a study of the frequency distribution of voice constructions in Chamorro and comes to the conclusion that active-direct clauses are most frequent, and that inverse, passive and anti-passive follow (in this order), with a much lower frequency each - a cross-linguistic distributional profile. Although Givón admits that 'frequencies vary with text-type, genre, author and more subtle sub-functions', this variation is not taken into account when the data are compared: a simple and straightforward comparison of frequencies (in %) is all.

When the variation Givón mentions had been taken into account, his conclusion could have been different. In particular, it could have been the case that the difference in frequencies were smaller. This is because the differences found may have been at least partly attributable to the variation between text-type, genre, author, and the more subtle sub-functions Givón mentions, and not to the distribution of voice constructions independently from those other factors. The differences in frequencies, therefore, attested here as a cross-linguistic tendency, may have been 'explained away', so to speak, by other factors causing the differences.

In this paper, we will consider the consequences of taking those factors into consideration that are not usually considered to be important, but that may cause frequencies of linguistic constructions to vary. In particular, we will look at the occurrence of the passive voice construction in Dutch texts. It is often assumed that, for example, the passive is less frequent (than average) in spoken language, and more frequent in the government and policy texts (see, for example, Vandenbosch 1992 and Renkema 1981). This conclusion is drawn on the basis of, again, frequency data of the specific genre, this time without taking the influence of individual texts, authors (and perhaps Givón's 'more subtle sub-functions') into consideration. However, passives do not just occur in genres; they - first and foremost - occur in *texts*. There is a nesting: passive clauses occur in sentences that occur in paragraphs that occur in texts that occur in a certain genre. At which level of this nesting is the attested variation demonstrable? It has always been assumed that it is at the level of genre, but are we really sure about that?

What we are actually going to show can be explained by means of the following non-linguistic example. Imagine the situation in which a researcher studies the time spent watching television in families. He finds a family with eight members (two adults and six children) that do not watch television at all (0 hours), and he finds a family of two members that watches 2 hours of television a week. What is the average? In effect a two stage sampling procedure is in operation. In the first stage households are randomly selected from a population of households. In the second stage individuals within a household are drawn. In accordance with the sampling procedure a distinction can be made between the variance *within* primary units (households in this example, or texts in a corpus) and variance *between* primary units. The first variance component is an estimate of the differences within primary units (variance between persons within a household, variance between sentences within texts), and the second is indicative for the differences between primary units (variance between households, and variance between texts). We cannot consider each element within a primary unit as completely independent from other elements in the same primary unit. Persons within a household negotiate - to some extent - whether they will watch television; children are sometimes allowed, and sometimes not, to watch television. Dependent on the beliefs of parents children are allowed to watch more of less often.

The same holds for sentences within a text. Two sentences drawn aselect from one text are more alike than two a-select sentences drawn from different texts. First off all these sentences deal with the same topic, and are likely produced by the same author (or team of authors). If each element is considered as an independent observation we clearly neglect this information. However, as each element within a primary unit is related to other elements in that primary unit (the dependency), we use the information which was shared by the elements within that primary unit more than once. This results in an overoptimistic view on the data. That is, testing statistics, like χ^2, t, or F, are inflated to an unknown degree, and standard errors are too small. This is merely another way of saying that cluster samples are less precise compared to single samples of the same size (compare, Kish 1965; Cochran 1979). In the television example the population mean is estimated as the mean of household means (i.e. one hour), and not as the mean of the individual observations (i.e. 12 minutes).

When this line of reasoning is applied to the occurrence and frequency of the passive in Dutch texts, the following holds. On the theoretical side, there is evidence that the occurrence of a passive may make it more likely that another passive occurs: in production, one passive may 'prime' another one. This dependency has been attested for spoken language in the literature by means of experiments, for example in Bock (1986) for English, and in Hartsuiker (1996) for Dutch (under some conditions). It is our impression that in texts, passives tend to occur together, in 'clusters'. In Cornelis (1997), it is shown that although each individual passive contributes a 'meaning' to the representation of the text, it is only when considered together that it becomes possible to consider the *effect* of the passive in a certain text: passives *together* function in a certain way. Therefore, there are plenty of reasons to consider the passives in texts dependent phenomena: dependent on the *text* they occur in, not only the genre. However, this acknowledgement of the dependency of occurrence has not found its way to the quantitative methods used by text linguists. Therefore, the methodological/statistical side to the question is still, indeed, an open question. In this paper, we will propose an answer.

2. The analysis: the occurrence of the passive in a corpus of Dutch texts

For our analysis, we used the corpus described in Vandenbosch (1992), a corpus of Dutch consisting of entire texts and large text fragments, both from the Netherlands and Belgium. The corpus consists of two large subcorpora, written language versus spoken language, each 50 % of the entire corpus. The written language subcorpus consists of two genres, argumentative &

popular-scientific texts ('arg-popular' from here on) and narrative/fiction ('narrative'). Vandenbosch (1992:55/56) assumes the passive to be most frequent in arg-popular texts (13.4 %), and least in narrative (2.9 %), with the spoken language holding a middle ranking (3.4 %).

We considered all texts in each each of Vandenbosch's three genres ('text' therefore also refers to the fragments in the spoken corpus). Of each text, we took the middle page (retrieving Vandenbosch's corpus in WordPerfect 5.1, standard page lay-out, font times 10), we took the middle page in case of an odd number of pages, and the page number as the total number divided by two of with an even number of pages (i.e. of a text of 12 pages, the sample page was 6, of a text of 15 pages, it was 8). The length of the fragments in the Vandenbosch corpus varies from 3 to 44 pages, with an average of 16.2 pages for the written corpus, and 9.3 pages for the spoken one. Of these fragments, we determined the number of clauses and whether that clause was passive or not-passive. Our definition of clause includes non-finite clauses, such as *de affaire is kennelijk te controversieel om te worden behandeld*, ('the affair apparently is too controversial to be dealt with') but not verbless utterances (*o, ja* 'oh, yes'). In general, we followed Vandenbosch's transcription, considering capital letters to demarcate sentences.

'Passive' was defined as any construction of *worden* (literally: 'to become') and a past participle, following Cornelis (1997), who argues that only the construction with *worden* is the real Dutch passive. This means that constructions of *zijn* ('to be') with a past participle, often considered to be the 'perfect aspect' of the Dutch passive, as well as past participle constructions with other *be*-like auxiliary verbs were left out of consideration (see Cornelis 1997 for a motivation). This means that 'non-passives' does not equal 'active' in our analysis. We will therefore use 'non-passive' versus 'passive'.

Summing up: every observation was coded for genre (3 levels: spoken language, arg-popular, narrative), text number, sentence number, clause number, and passive versus non-passive.

A standard analysis can be performed by either a χ2-test, a logit analysis, or even a oneway analysis of variance (compare Feinberg 1980). The data concerning both analyses are summarized in Table 1.

Table 1 Frequencies of passives per genre and mean number of passives (between brackets)

	Arg- popular	Narrative	Spoken	Total
Passive	38 (.11)	3 (.01)	16 (.03)	57 (.05)
Not passive	303	264	545	1112
Total	341	267	567	1169

According to a traditional χ^2-test the null hypothesis of equal proportions of passives has to be rejected ($\chi^2 = 41.94$; d =2; $p < .001$); the number of passives is not equally distributed across the three genres. Both a logit analysis and an analysis of variance lead to the same conclusion; the null-hypothesis (of equal cell means) has to be rejected (respective testing statistics are: $G^2 = 38.92$; df = 2; $p < .001$, and F= 21,70; df = 2, 1166; $p < .001$). Hence, no matter the type of analysis, we have to reject the null-hypothesis (of equal frequencies, proportions, or cell means) for the three different genres.

Next we turn to a multilevel model. In the data clauses are coded as being passive. Clauses are nested within sentences, and sentences are nested within texts. Hence, we have three variance components, the variance between clauses within sentences, the variance of sentences within texts

and the variance between texts, and three means, one for each genre. Note however, that the variance of clauses within sentences is assumed to be binomially distributed, and hence is a function of the mean of genre ($\sigma^2 = p * [1 - p]$). Suppose, Y_{ijk} is the observed score for the *i-th* clause (coded 1 for passives, and 0 otherwise) of the *j-th* sentence in the *k-th* text, the model to be analyzed can be written as:

$$\text{Logit}\,(P_{ijk}) = G1_{ijk}\,[\,\beta_1 + u_{10jk} + v_{100k}] + G2_{ijk}\,[\,\beta_2 + u_{20jk} + v_{200k}] + G3_{ijk}\,[\,\beta_3 + u_{30jk} + v_{300k}]. \qquad (1)$$

In equation (1) $G1_{ijk}$, $G2_{ijk}$ and $G3_{ijk}$ are dummy variables which indicate the three genres. That is $G1_{ijk}$ equals 1 only if an observation was in an arg-popular text, otherwise $G1_{ijk}$ was coded zero. The same holds for $G2_{ijk}$ and $G3_{ijk}$ ($G2_{ijk}$ equals 1 only if a clauses was observed in a narrative text, and $G3_{ijk}$ equals 1 only if a clause was observed in a spoken text). Per genre a mean is estimated (β_1, β_2 and β_3). Furthermore, two residual scores are estimated per genre: one residual for the *j-th* sentence (u_{10jk}, u_{20jk} and u_{30jk}) and one for the *k-th* text (v_{100k}, v_{200k} and v_{300k}). The last residuals represent the deviation for the *k-th* text of the estimated mean score. Hence, the mean of text *k* in the first genre equals $\beta_1 + v_{100k}$. The first mentioned residuals, the residuals at sentence level represent the deviation of the *j-th* sentence for the *k-th* text. Hence, the mean number of passives in the *j-th* sentence of the *k-th* arg-popular text equals $\beta_1 + v_{100k} + u_{10jk}$. The residuals are assumed to be normally distributed (which is not a strange assumption, as it concerns a distribution of means).

The model was estimated on the same data as presented in Table 1, which were used for three types of unilevel analyses. The parameter estimates are summarised in Table 2.

Table 2 Parameter estimates according to a multilevel model (see equation 1; standard errors between brackets).

Parameter	Arg-popular	Narrative	Spoken
β (logit)	-2.07 (0.58)	-3.66 (0.58)	-3.53 (1.52)
Proportion[1]	0.112	0.025	0.029
Variance between texts	0.62 (.14)*	0.00 (0.00)	0.00 (0.00)
Variance between sentences	0.00 (0.00)	41.03 (5.30)	0.00 (0.00)

1: remember: proportion = ln (β) / [1 + ln (β)].

To test the differences in means a contrast analysis can be performed (e.g. Goldstein 1995; Bryk & Raudenbush, 1994). This yields an overall result of $\chi^2 = 3.94$; df = 3; $p > ,05$. Hence, we cannot reject the null-hypothesis that all three means are drawn from the same population, or to put it differently, we cannot show that the three means are different.

It cannot be concluded however that arg-popular, narrative and spoken texts are comparable with respect to the number of passives. We also have to take the variance estimates between texts as well as the variance estimates between sentences into account. Table 2 shows that arg-popular texts differ from zero (i.e. the parameter estimates exceeds twice the standard error). Hence, the number of passives; the number of passives varies from text to text; the number of passives is relatively large in one arg-popular text but small in another. We can conclude that some authors use passives more often than others. This is not surprising when we take into account that Vandenbosch's category of argumentative and popular-scientific texts is rather broad: it includes for example a report of an archaeological expedition, but also a policy statement of a broadcasting company.

For neither narrative or spoken texts differences between texts could be shown. For narrative texts however, there clearly is a variance between sentences. Hence, there is a clustering of passives within texts. The data suggest that for narratives one passive clause within a sentence triggers a passive in the next. It is not clear exactly what causes this triggering. It seems likely that some perspective phenomena may be the cause, but it remains an issue for further research to investigate why it only occurs in the category of narrative texts and not in the others. We see, however, that only a multilevel analysis shows that the clustering exists.

3. Conclusion

In contrast to a more traditional analysis, a multilevel model analysis of frequency data of the passive in a corpus shows that the attested difference in proportions of passives between genres has to be attributed to a difference between texts. When this variance is taken into account, there is no significant difference in the proportions of passives between genres. There are, however, other interesting differences that only appear when these multilevel phenomena are analyzed in a multilevel analysis: the variance between texts is greater in argumentative and popular-scientific texts than in the other genres (narrative and spoken language), and in narrative, there is variance between sentences that so far, the theory cannot account for. The implication of this is that whereas a unilevel analysis of multilevel phenomena can lead to serious inferential errors (cf. Cronbach 1976), a multilevel analysis of those phenomena may open the way for new theory-building.

4. Literature

Bock, J. Kathryn (1986) 'Syntactic persistence in language production'. In: *Cognitive Psychology* 18, 355-387.

Bryk A.S. & Raudenbush, S.W. (1992) *Hierarchical linear models: Application and data analysis methods*. Newburry Park: Sage.

Cochran, G. (1977) *Sampling techniques*. New York: Wiley.

Cornelis, Louise H. (1997) *Passive and perspective*. Amsterdam/Atlanta: Rodopi. Utrecht Studies in Language and Communication 10.

Cronbach, L.J. (1976) *Research in classrooms and schools: Formulations of quistions designs and analysis*. Occasional paper, Stanford evaluation consortium.

Feinberg, S.E. (1980) *The analysis of cross-classifiek categorical data*. Cambridge: MIT press.

Givón, T. (1993) 'The pragmatics of voice: functional and typological aspects'. Reader 2, IFOTT lectures, Amsterdam, may 24-28.

Goldstein, H. (1995) *Multilevel statistical models*. London: Deward Arnold.

Hartsuiker, Robert J. (1996) *Sentence Production in Normals and Broca's Aphasics: Stages and Resources*. Dissertation University of Nijmegen.

Kish, L. (1965) *Survey sampling*. New York: Wiley

Renkema, Jan (1981) *De taal van 'Den Haag'. Een kwantitatief-stilistisch onderzoek naar aanleiding van oordelen over taalgebruik.* 's-Gravenhage: Staatsuitgeverij.

Vandenbosch, Luc (1992) *Aspekten van passiefvorming in het Nederlands. Een kognitief-pragmatische benadering*. Dissertation, University of Antwerp (UIA).

Quantity and style from a cognitive point of view

Gábor Tolcsvai Nagy

Address: Soumalais-ugrilainen laitos
PL 25 (Franzeninkatu 13) FIN - 00014 Helsingin Yliopisto
Fax: (3586) 1917019
E-mail: tolcsvai@cc.Helsinki.fi
Affiliation: guest associate professor of Hungarian at Suomalais-ugrilainen laitos (Finno-ugric Department) Helsinki University

Summary: Quantity as a source of style has not been yet interpreted in a pragmatic and cognitive frame. The theoretical part of the paper gives a summary about both the traditional and the pragmatic style theories, concentrating on the question of quantity. The examples show the possibilities of cognitive analysis and give models of perception, ways of cognition and understanding concerning quantity and frequency.

Topical paper

Topic area: stylistics, model construction, explanation of text phenomena

Theoretical issues

Most traditional style theories are based on a kind of descriptive grammar, on a grammar of rules and lexicon and on a grammar of standard language. These style theories take Saussure's *langue* as the basis of the explanation: speakers of a language can possess language knowledge (i. e. the knowledge of the *langue*, the homogeneous and common system of language) in the same way and manner. Therefore a text can be received and understood by many hearers/readers with the same processes. In these theoretical frames features of the text seem to be immanent characteristics, approachable by different hearers/readers in the same way. Within these theories quantity is not a problem: quantity can be measured, it can be explicitly analized, so it has a big explanatory power concerning both the process of understanding and the resulting effect (*Wirkung*). Quantity therefore can be described in the frame of a grammar, by counting the frequency of certain discrete linguistic elements within the text, and comparing these data with the average of the same (standard) grammar. In one of the main trends of these style theories style arises when the text data show some differences compared to that of the grammar, thus style is a result of deviation (*écart*; cf. Guiraud 1970). (Of course deviation of quantity is only one – important – component of the deviation theories.) The other trend considered style as a result of selection (of the elements from the *langue*) and the combination (of the selected elements) (Cressot 1947, Marouzeau 1949; Jakobson 1960 in a different frame). The quantity of linguistic elements – as part of the style of the text – is also the result of selection.

Recent style theories question the central role of grammar, more precisely the role of autonomous syntax suggested by the classical generative theory. Instead these theories approach style as a hermeneutic or semiotic phenomenon: "Das stilistische Zeichen aber ist nicht vor oder jenseits seiner Bedeutung schon als Zeichen gegeben, und stilistische Zeichen lassen sich nicht wie Wörter, unabhängig vom Text, den sie kennzeichnen, inventarisieren. [...] *Die Merkmalsbestimmung ist von der Interpretation des Merkmals nicht zu trennen*" (Anderegg 1977: 57; see also Anderegg 1995, especially 123). This feature comes from the assertion that style is a textual phenomenon, it can be understood only in given circumstances. As text is not purely an entity of grammar but it has a pragmatic character (action, situation and context as components of the verbal interaction take part of its formation and understanding), style is to be explained in a pragmatical view of language, too. These ideas are explicated for instance in the works of M. A. K. Halliday (1978), N. E. Enkvist (1978), J. Anderegg (1977), B. Sandig (1986); see also Spillner (Hg. 1984), Gumbrecht – Pfeiffer (Hg. 1986), Stickel (Hg. 1995). The final conclusion of these ideas is that style is part of the sense (*Sinn*) of the text, but it does not come from the cognitive meaning of the words, expressions, sentences of the text, but from its formation. In the context of this paper formation means the process of giving a text a certain form, and also the result of this process, the "form" of the text, wich means a widely understood language formation from phonology to syntax in a pragmatic frame. Sense is used here for the "meaning", "*Sinn*" of the text, according to the hermeneutic discretion (cf. the same standpoint in Beaugrande – Dressler 1981: 89–91)

Style as formation can be produced and recognized in specific socio-cultural circumstances: the speaker or the hearer have to know something about the style types of his/her language community. On the basis of this knowledge the speaker has a certain target norm (*Zielnorm*) and effect intention (*Wirkungsabsicht*), the hearer has an expectation norm (*Erwartungsnorm*). This knowledge of the speaker and the hearer comes from the experiences of everyday verbal interactions, it is organized similar to other kinds of cognitive abilities, and it can be regarded as the general base (Langacker 1987) or context (Givón 1989) that gets into relation with the text in question by the process of comparison. The speaker or the hearer always considers the style of a given text as the token of a style type (like other linguistic expressions as the token of certain semantic, syntactic or phonological types; cf. Langacker 1987). This classifying procedure takes place according to the prototype theory (Rosch 1977). Since the speakers of a language community do not know the same types and sometimes they classify the same texts in different types (according to their socioregional origin and language knowledge), the hermeneutic characteristic of style is explicated on the global level of understanding first (I give a detailed explanation in Tolcsvai Nagy 1996).

Does quantity have any kind of meaning in this approach of style? Does it have such significance as it had in the traditional ones? In most of the above cited works quantity is not even mentioned. Although Enkvist (1978) considers it as an important part of his pragmatic style interpretation.

The first point of the answer is that the hearer or the reader does not count the discrete elements of a text. He/she makes comparisons (one of the basic processes of cognition) between 1) the relative frequency or quantity of one element (or more elements) of the text and the other elements with neutral frequency in the text; 2) the relative frequency or quantity of one element (or more elements) of the text and his expectations based on the assumed text type and the style type to which the text belongs

in his/her judgement. These comparisons are made during the reception of the text, and quantity means here not exact measurements but rather approximations based on the prominent elements of the text. The prominent elements are those that become style elements for the hearer/reader during the first hearing/reading (style element is that linguistic expression, which has a certain style value, i. e. in a certain text it belongs to a certain style type for some reasons for a speaker or a hearer), and they can appear at every linguistic level (phonetic/phonological, morphological, syntactic, semantic, lexicon level), having validity always at the textual level. Prominency is determined by the interaction of the reader/hearer's expectations and the recognized text elements. Effect (*Wirkung*) is therefore based on the cognition of certain elements by the first, spontaneous hearing/reading. This first reception can be followed by a second one that is based on the experience of the first one, and this second hearing/reading can explain the style features of the text in a more explicit conceptual frame and can recognize better the quantity characteristics of it, too (cf. Jauß 1991: 813–46).

In this frame quantity has significance, but it becomes a relative feature: frequency or quantity expressed by objective numbers in a description may have different effects on different hearers/readers or at the extreme may have no effect at all. The relative characteristic is of course restricted: a culture or a language community have some kind of *sensus communis* concerning the relation of text type, style type (and action, situation, context) and the quantity of certain linguistic phenomena (e. g. sentence length or sentence complexity), but nevertheless dispersion may be large.

Example 1.

The first example of this paper concentrates on the relative frequency of certain elements comparing them with other ones within one text. The text to be analyzed here is *The Love Song of J. Alfred Prufrock* by T. S. Eliot. One of the most fascinating feature of Eliot's poetry is the presence and effect of objective correlatives, as the poet himself consciously constructed his objective lyrics with the help of these elements. There are many objective correlatives in *Prufrock*, namely 53 objects are mentioned as part of the scene and part of the emotive structure of the poem. The analysis is going to point out the relative frequency of these elements, the repeated parts where the frequency is higher, and where there is no objective correlative. The comparing activity of the reader remains mainly within the text. The main concern of the present analysis is nevertheless the deeper sense of this wave-like frequency. The objective correlatives can be classsified semantically in the following groups: urban street scene, five o' clock tea, human body, clothes. All correlatives are prototypes of their own type and all are basic level categories (between superordinate and subordinate categories; cf. Rosch 1977, Lakoff 1987: 46pp). The frequency and sequence of these elements in the abstract intellectual context of the poem creates a special duality, and it is of course prominent enough for the reader. The other main question in *Prufrock* is the semantic range of the objective correlatives. A thorough analysis has to point out that in some parts where frequency is higher, the semantic range of the correlatives contact each other. At other parts of the poem it is on the contrary, although there are repeated correlatives (e. g. ***street***). The correlatives are cognitive units (cf. Langacker 1987: 57pp), but their semantic ranges change according to their frequency and joint quantity.

Example 2.

The other example to explain here is the sentence length and sentence complexity of *Dubliners* and *Ulysses* by James Joyce. Opposite to example 1. here the comparing activity of the reader may concern the inner characteristics of the texts and the reader's language knowledge of the relation between sentence type and prose. Quantity (and not frequency) is the main question. In *Dubliners* the average sentence length is quite short and the average complexity is relatively simple, morover the whole volume is homogeneous in this respect. The reader experiences this homogeneity and may judge the sentence length in different degrees short and simple. The analysis has to point out what kind of unit (or units) can be perceived in the text by the reader as sentence patterns (*Gestalts*) on the basis of length and complexity and what are the possibly nearest other units outside the text as the other side of the comparison. With *Ulysses* the case is different: in this novel there are different types of sentences, short and simple, short and complex, long and complex, so the comparing activity of the reader remains to a larger extent within the text, although *Ulysses* is much more a provocating novel than the volume of the early short stories. Quantity gets (or may get) its validity in *Ulysses* therefore considerably within the text.

In both examples the analysis has to show the role of iconicity (especially diagrammatic iconicity) in the possible ways of understanding.

Conclusion

Quantity and frequency as a source of style can be interpreted also in a pragmatic and cognitive frame. The theoretical part of the paper gives a summary about style theories of this kind, the examples show the possibilities of analysis in this frame. The analysis as well as the theoretical approach have to give models of perception, ways of cognition and understanding concerning quantity and frequency as one possible way to form style in a text.

References

Anderegg, Johannes 1977. *Literaturwissenschaftliche Stiltheorie*. Vandenhoeck & Ruprecht. Göttingen.

Anderegg, Johannes 1995. Stil und Stilbegriff in der neueren Literaturwissenschaft. In: G. Stickel (Hg:): *Stilfragen*. Walter de Gruyter. Berlin, New York. 115–27.

de Beaugrande, Robert-Alain – Dressler, Wolfgang U. 1981. *Einführung in die Textlinguistik*. Niemeyer. Tübingen.

Cressot, Marcel 1947. *Le style et ses techniques*. Paris.

Enkvist, Nils Erik 1978. Stylistics and Text Linguistics. In: W. U. Dressler (ed.): *Current Trends in Textlinguistics*. Walter de Gruyter. Berlin, New York. 174–90.

Givón, Talmy 1989. Mind, Code and Context. Essays in Pragmatics. Lawrence Erlbaum Associates Publishers. Hillsdale, New Jersey, London.

Gumbrecht, Hans Ulrich – Pfeiffer, Ludwig K. (Hg.) 1986. *Stil. Geschichten und Funktionen eines kulturwissenschaftlichen Diskurselements*. Suhrkamp. Frankfurt.

Guiraud, Paul 1970. *Problemes et méthodes de la stilistique*. Paris.

Halliday, M. A. K. 1978. *Language as a Social Semiotic. The Social Interpretation of Language and Meaning*. University Park Press. London, Baltimore.

Jakobson, Roman 1960. Closing statement: Linguistics and Poetics. In: Sebeok, T. A. (ed.): *Style in Language*. Cambridge, Massachussettes. 350–77.

Jauß, Hans Robert 1991. *Ästhetische Erfahrung und literarische Hermeneutik*. Suhrkamp Taschenbuch. Frankfurt am Main

Lakoff, George 1987. *Women, Fire, and Dangerous Things*. The University of Chicago Press.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar. Volume I*. Stanford, California.

Marouzeau, Jean 1949. *Précis de stylistique française*. Paris.

Rosch, Eleanor 1977. Human Categorization. In: N. Warren (ed.): *Studies in Cross-Cultural Psychology*. Academic Press. London. Vol. I. 1–49.

Sandig, Barbara 1986. *Stilistik der deutschen Sprache*. Walter de Gruyter. Berlin, New York.

Spillner, Bernd (Hg.) 1984. *Methoden der Stilanalyse*. Narr. Tübingen.

Stickel, Gerhard (Hg.) 1995. *Stilfragen*. Institut für Deutsche Sprache Jahrbuch 1994. Walter de Gruyter. Berlin, New York.

Tolcsvai Nagy, Gábor 1996. *A magyar nyelv stilisztikája*. [The stylistics of Hungarian.] Nemzeti Tankönyvkiadó.

LOGICO-SEMANTIC AND STATISTICS APPLIED TO TEXTUAL DATA IN INFORMATION RETRIEVAL

Omar LAROUK

École Nationale Supérieure des Sciences de l'Information et des Bibliothèques (CERSI)

17-21, Boulevard du 11 Novembre 1918. 69623 VILLEURBANNE Cedex , FRANCE

Phone: (+33) 72. 44. 43. 43 Fax: (+33) 72. 44. 27. 88 E_mail: larouk@enssibhp.enssib.fr

Abstract:

In this article, we describe the problem of the conjunctions of coordinations in the French language.This paper treats of the textual data. Most document retrieval that user queries be specified in the form of boolean expressions. The are the uncertainty in combining queries, They have flaws. Many ambiguities in texts are due to the use of classical methods of computing. The idea of the work is to extract the "reach" of the coordination (conjunctions) in information system and looking for the importance of punctuation with statistical. Textual algoritm is contributed to detection and correction the signs of punctuation.

Keywords:

Uncertainty textual data. Connectors. Logics. Computational linguistic. Algorithm of detection and correction. Statistics for punctuation of signs. Quantitative Linguistics. Documentation and IR.

1. Introduction

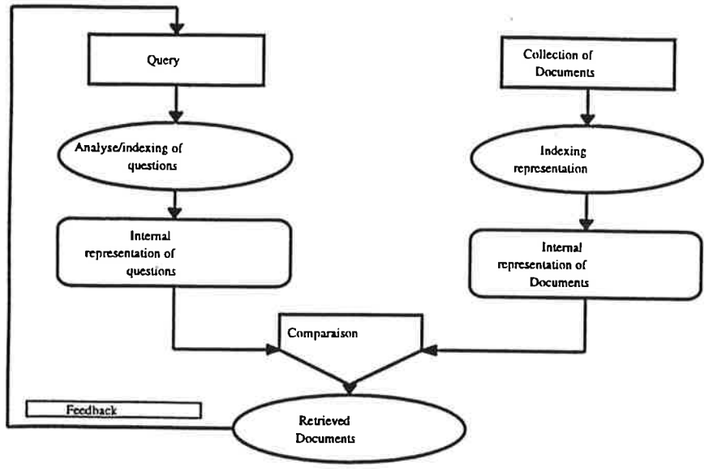
In this paper, we will describe an approach of automatic analysis of written *french* texts. This analysis is based on a linguistic model to determine informational elements of text. We will show that this type of analysis can operate with lexicon. Systems of automatic natural language analysis work on the bases of the principle of the word recognition aided with the list of textual form (free predicates) in lexicon.

The aim of the project is to design a DIS using natural language. Knowledge representation is a crucial point in Expert Systems technology and linguistics is at the crossroads of Artificial Intelligence [12]. Documentary automation is blocked by the problem of indexing and interrogation due to the constant updating of the input information which remains static once in the data base.

2. Text segmentation in Texts Information Systems

2.1. Documentary Information Systems (DIS)

An DIS is a system which can be used to manage large archives of documents and allows the user to store and to retrieve document which satisfy the questions. The general model of Information Retrieval Systems (IRS) [9] is given in the next figure.



The entities are :

- a collection of documents to indexing,
- The user formulates his request in a formal query language,
- a comparison identification.

The documents are indexed and stored. The user is formulised his query in a the information language who the system can understand: *request language [4]*. The query is compared against the documents. For the comparison to be effective, the query of the user must be translated into the indexing language. The comparison operation of the query with the representation of all the documents. When the documents are selected and which satisfy the questions according with the model. The user evalutes the documents according to his information needs. If the user is not completely satisfied, he can reformulate his request to system [10].

Therefore the idea was put forward to use natural language as textual data for automatic indexing because language integrates the temporal factor (linguistcal data: through connectors, verbs, adverbs, etc.). Therefore the priority of the group was to design a morpho-syntactic model and statiscal tools of written text (in this case *french*).

2.2. Text segmentation in documentation

Text retrieval systems are generally built on the reductionist basis that words in texts (keywords) are used as indexing terms to represent the texts. A necessary precursor to these systems is "word extraction" which for *english* texts, can be achieved automatically by using spaces and punctuations as word delimiters. This cannot be readily applied to *french* texts because. A lot of problems include segmentation ambiguity are in classical models of decomposition.

Textual analysis may be defined as the segmentation of texts into linguistic units, normally *words*. It is a precusor to text retrieval and generally to natural language processing systems. To take french text retrieval systems as an example, texts are segmented into words using spaces and punctuations as word delimeters; these words (or some of them) can then be used for indexing and retrieval.

3. Uncertainty problem in Information System

3.1. Boolean connectors and queries

A set *(the collection)* **D** of documents exits for retrieval. A set **I** of index terms *(the indexing vocabulary)* exits also with a function f mapping ***DxI*** into **{0,1}**:

$f_{(d,r)} = 1$ if document *d* is "implies" *r*
$= 0$ if document *d* is " not implies" *r*

In response to the simple query "r_0" :

(i.e. "r_0" ="*Give me all the documents 'implies' r*),

a retrieval system would retrieve for a user the following set :

$$\Re \{ d \in D : f(d, r) = 1 \}.$$

More complex queries can be constructed by allowing the ordinary boulean operators AND, OR, and NOT to connect terms to form expressions, one example being :

((p AND q) OR t) AND (NOT (z OR y))

We can define sets (P, Q, T, Z, Y) that correspond to boolean operations :

((p AND q) OR t) AND (NOT (z OR y))

$$((P \cap Q) \cup T) \cap (\neg (Z \cup Y))$$

(p **AND** q) defines $(P \cap Q)$

(p , q **AND** t)=(p **AND** q **AND** t) defines

$$(P \cap Q \cap T)$$

It's possible to calculate with the classical logic.We have again used the Boolean truth tables:

a	*b*	*a* AND *b*	*a* OR *b*	NOT *a*
0	0	0	0	1
0	1	0	1	1
1	0	0	1	0
1	1	1	1	0

We recall that the common functions used to produce these tables are:

a AND b	is either	a*b	or Min(a, b)
a OR b	is either	a + b - a*b	or Max(a, b)
NOT a	is	1 - a	

3.2. Families of ambiguities related to linguistical connectors

We shall see to what extent the connectors in the query give rise to ambiguities [6]. We shall present examples illustrating different meanings of the connector "AND" which is considered as a syntactic operator of the intersection.

Addition
<a> *Paul likes fruit **and** cakes*

Succession
<b> *The robber hit me **and** he ran away*

Combination
<c> *The material is black **and** white*

Presupposition
<d> *The guard saw the fire **and** gave the alert*

Elipse
<e> *The boy bought a shirt **and** so did the girl*
i.e. *the boy bought a shirt and the girl bought a shirt too*

The connector "AND" has a single meaning in formal language (reductional), which is not the case in Natural Language [5].

3.3. Criticism of the purely computerized system : Eliminates the connector « COMMA »

Many ambiguities in texts are due to the use of classical methods (boolean logic). In Boolean logic recognition, the meaning of conjuctions in natural language is ambiguous. For example:

*i) Natural language understanding **and** logic*
*j) Logic , uncertainty **and** retrieval information*

This sentence could be interpreted as :

i) Natural AND language AND understanding AND logic

*j) Logic **AND** uncertainty **AND** retrieval **AND** information*

A different problem occurs when the operator of negation is used. In linguistical analysis, the operator *"NOT"* is often emphasized as in **"BUT_NOT"** or **"AND_NOT"** [5,7]. A different approach is proposed by Salton [9]. He propose to implied boolean logic by taking identified terms of a query and submitting an initial search, using only the connector OR, expecting to retrieve a large number of items. Das-Gupta [2] developed a procedure to identify the Boolean operators AND and OR in natural language queries using syntactic and semantic information about conjunctions.

The classical model eliminates words such as: (/AND/, /as/, /but/, /comma/, /semi colon/, etc...) which serve as indicators for dividing the sentences into propositions, these words being considered as "empty". However, the sentence is recognized as being the largest linguistic entity and the coordination is considered as the process for forming new sentences following the sequencing law such as :

$$P = \Sigma P_i = P_0 + P_1 + \ldots + P_i + \ldots + P_n$$

where the connector "AND" can replace the sign "+" as a means of building the structure, thus resulting in an information loss.

The different linguistic [9] causes have the effect of reducing the level of relevance by selecting documents which are not required. At the implementation stage (document search), the analysis of the query must be considered. Therefore it is during the modeling phase of the information stock documents that the effectiveness of a DIS can be determined. This first phase is closely related to textual element processing, including the words considered by certain computer experts as being "empty" such as connectors, prepositions, etc...

3.4. Fuzzy logic applied to connectors

Zadeh [14] developed a new approach to represent intermediate values between 'true' and 'false' using an extension of the model of Lukasiewicz. This approach is known as Fuzzy Logic or Imprecise Reasoning. Fuzzy logic is a generalization of the mathematical notion of set membership, in which an element may have partial membership in a set. Zadeh allowed for an infinite range of values between 0 and 1 whereas the classic logic allowed 'false and 'true' [3], [13], [14], [15]. He used the connectors "AND", "OR" with classic formalism of set, but in the place of the addition (OR), he used the MAX for maximum and instead of using multiplication (AND), he used MIN for mininum. For example, in fuzzy logic, we treat the problem of connectors with the following example :

<j> /Paul is very tall AND Paul is very smart/
A C B

Further more, let us assume that */"Paul is very tall"/* which will define (with a probability) A= 80% to reflect "very tall" and */"Paul is very smart"/* which will define B=80% to reflect "very smart". The following statements demonstrate the difference between Extensionnal Logic (classic logic) and Fuzzy Logic (non-classical logic) :

A= "Paul is very tall" = 80%
B= "Paul is very smart = 80%

This technic of probability is linked to user who treating this information. So, we will obtain :

Extensionnal Logic	: A AND B
Operation	: 80% * 80%=64%
Fuzzy Logic	: Min [80%, 80%]= 80%

The analysis of the connector "OR" in the following statement (proposition). By appling the operation of distributiveness in this example, we will obtain :

<jj> /Paul is tall OR Paul is smart/
A C B
/Paul is tall OR smart/
/Paul is tall OR smart/

The combining of predicates (adjectives) is not correct

<iii> /Paul is very tall OR Paul is very smart/
A C B

Extensionnal Logi	: A OR B
Operation	: 80% + 80%=160%= 1.6 (converted to 1)
Fuzzy Logic	: Max [80%, 80%]= 80%

The result produced (64%) in the case of the connector "AND" is lower than the result produced be the non-classical logics (Fuzzy logic). If we thinks that 64% represents the prédicate "quite tall" (*plûtot grand*) whereas 80 % represents "very tall" (*très grand*). We constates that the fuzzy logic isnot "optimal" for the analysis of connectors.

We propose to "optimize" information search through text processing and query formulation in natural language using the contribution of logico-semantic approach applied to connectors.

4. Logico-semantic applied to the connectors and punctuation

The punctuation is a means to divide the parts in a speech. Let's see had the signs of the connective punctuation act in a sentence. Indeed, the connector "Comma" opposes itself to the connector "and" and points out the enumeration. The last "and" announces the closing of the chain. Let the following example :

*<1> /the enterprise gives information on the structures , the posts **and** the persons./*

*<1> /The enterprise gives information on the structures **and** the posts **and** the persons/.*

The open chain (without a final connector) shows the non of an enumeration.

<2> /The enterprise gives information on the structures, the posts, the persons./

The open chain <2> presents a strict distribution such as :

S1. -"The enterprise gives information on the structures"
S2. -"The enterprise gives information on the posts"
S3. -"The enterprise gives information on the persons".

On the contrary, the closes chain <1> presents **a total distribution** (four solutions : S1, S2, S3, and S4 linked to the combinative connector).

S4. -"The enterprise gives information on the structures **and** the posts **and** the persons".

This calculation of the Logico-Semantic-Images (LSI) allows to calculate every extensions of an ensembiste text. We get a referential calculation of each part of the textual data.

5. Statistics analysis of the textual data

We're going to try to see the links of the signs of punctuation with the connectors. We have noted that the punctuation of text is richer than this of the sentence. These different signs of punctuation are :

Comma, point_comma, suspension, point, capital_letter, full_punt, explainative_dash, cupple of dashes, colons, quotations, cupple of brakets, square_brakets, exclamation_mark, interrogation_mark,...

We'll use technics of the statistics to describe the textual information in order to find a sign of punctuation. And see which is its real coordinative value ?

5.1 Distribution of the textuals signs

The linguistic predicates such as compounds and phrases are not cleary definable. for example, is breakfast a simple word or compound? In information retrieval a compound or phrase? The fuzziness of these linguistic units

in french is somehow reflected by the fact that two formatives are sometimes written together, for example, *microordinateur* sometimes hyphenated *(micro-ordinateur)*.

A hierarchic classifying of the elements appeared is possible. The forms are in the next table :

CONNECTORS + Punctuations	Frequency	correction	real frequency
AND	5	+2	7
OR	0		0
COMMA_AND	1	+1	2
COMMA_OR	0		0
POINT_COMA_AND	0		0
POINT_COMA_OR	0		0
capital_letter	40		40
cupple_commas	0	6*2=12	6
comma	25	-12-2-1	10
point comma	0		0
suspension	0		0
point	12		12
full_punt	0		0
explainative_dash	1		1
cupple_of_dashes	10		10
colons; quotations	2		2
cupple_of_brakets	0		0
square_brakets	0		0
exclamation_mark	0		0
interrogation_mark	0		0
TOTAL	96		90

Numbre of comma = Simple_Comma_Numbre + Cuple_Comma

Numbre of comma= (10+2+1) + (6*2)=13+12=25

Icc= Indice of correction of the comma NO_connectors:

$$Icc = \frac{\text{Numbre of Comma - 2*Cuple\_Comma}}{\text{Numbre of Comma}}$$

$$Icc = \frac{(25-2*6)}{25} = \frac{25-12}{25} = \frac{13}{25} = 0{,}52$$

RIcc= Real Indice of correction of the comma WITH_connectors

$$RIcc = \frac{\text{Numbre of Comma - 2*Cuple\_Comma -Comma\_AND-Comma\_OR}}{\text{Numbre of Comma}}$$

$$RIcc = \frac{(25-2*6)-0-2}{25} = \frac{25-12-2}{25} = \frac{11}{25} = 0{,}44$$

5.2 Comment on the results obtained

The sum of the words of the corpus 1 is 436. So, il we want to knad the relations between the signs and the words, we only need to take the number off the frequencies of the capital letters. Then, (436/51 = 8,55) we have a sign of punctuation for 8,55 words.

The results confirm the variety of the natural language by the use of different sings. But, in the corpus (AFP), we have distributions "very close" from each other in the ten dispatches. Indeed, we note that the three dominant signs

(having a high relative frequency) are " the capital letter, ithe comma and the full point". The countage being made with a few linguistic hypothesis.

Concernant the others signs, distributions are different with a little and irregular appearence of signs like "the couple of inverted commas", "the couple of brackets", "the couple of dashes" and "the explainatory dash". At last, the signs that don't appear are : "semicolon", "suspension point", "colons", "square brakets", "exclamation mark", "question mark"...

The capital letter is the most important sign of this corpus (AFP). What conforms the hypotheses of the groupe CERSI that the (NP = Noun phrase) has extralinguistic reference in 40 far as the word beginning with a capital letter send either to a name (/"Paris"/) or a common name ("Capital"/)

We suppose that the textual linearity is ensured by the full-point between paragraphs to treat a corpus nowing several paragraphes like a set of sentences (in spite or the presence of the indented line). Then, this process will allow us not to "and" - beginning - of - sentence".

We noticed that :

- the links between a punctuation sign and the words of each corpus is at an average of 8.33 (on corpus of the A.F.P.),
- the dash placed out the beginning of the paragraph is explanatory,
- the couple of dashes of the structure (AFP) is assimilated to the brakets,
- the couple of dashes and the brackets act as a couple of appositive commas. Then, they are coordinating and explainative,
- the comma and the dash are specialized in the expression of the set which is, more or len, the logical equivalent of the coordination we saw that the enumeration of the predicats enter a "ensembliste" mould.
-We note that the dispatches treat many general informations : technical, political, ecomic, etc... That's not the case of books of literature.

5.3 Comments

Here the data that the tratment is not numerical but linguistic. We note that a "relative" regularity (uniformity) of three signs in the ten corpus. Indeed, the corrector "and" is shared "locally" in some paragraphs contrarily to the commas and to the capital letters that we find again distributed.

The maximale frequency is the biggest appearance of a date. The letter can be considered as a "discriminating" or "dominating" element. Comparing with the other data, besides the capital letter, the criters most discriminating are the commas and the correctors ("and", "comma_and") comparing with the other forms.

We note that the coordinations of predicats (Pi*) (« The black and red flag »)* are more important than the coordinations of SN *(« uremployment and prices increase »).* As well, the dashes, the brakets, the quotations and the cupple of commas have a dubble value : sometime an explainative value, sometime a mark of breaks of speeck introducing a clause.

Why do we interest ourselves in the sign of strong punctuation (ex : the full point). Simply because of the value of "Full_Point_And" in the treatment of written document. Indeed the informative value of "Full_Point-And" which is in fact the connector "Comma_And" 's.

5.4 The statistic results are they meaningful ?

In the analysis, we tried to lighten the particularities of the corpus (AFP). This statistic treatment between the analysis of the distribution of the signs of punctuation comparing with the connectors in order to find the coordinative commas from the neutra commas to give the comma its very value. The aim of this statistic estimation was to see the following points :

- The number of occurence of the connectors *(coordination, comma, ...)*,
- Their importance and their relations with the neighbourhood in order to determine the nature of the coordination of predicates, of the noun phrase or of sentences,
- Or don't corpus possess precise connectors ?

It seems that there is a predominance of the connectors "And" in the texts (AFP) comparing with the other conjuctions of coordination *(OR, NEITHER...NOR, BUT, SO, BECAUSE)*. So, the use of statistic calculation happens to be useful for the analysis of the comma in its links with the connectors (AND, OR).

The study of the punctuation shows its place in the written and spoken communications.
The signs of ponctuation belong to the syntaxe of the sentence and of the text.
Generally, these signs are not only limits of segments (cf. algorithm of Spang) but also signs that have "coordinative information).

We have seen that during information searching, data base interrogation is carried out with the aid of descriptors combined with connectors (*"AND", "OR", "EXCEPT"*). The following idea may be deduced : formal languages *seem* to be well defined semantically whereas natural language entails many ambiguities [5]. We found that the results were limited due to the preferential use of computer based solutions rather than a linguistic and statistics solution [6].

Why we interested to documentation and to linguistical expressions ? Because texts contain many punctuations and connectors.

Thus, if you compare different types of text corpora pertaining to different domains (interviews, political speeches, information system user queries...) you need information semantic. We analyse textual data (statistical studies on *Agence France Press* (AFP) despatches) and we propose algorithms for the automatic treatment of punctuation signs (determination of the value of the comma compared to other connecting words in a text string like ***"and", "or"***, ...).

6. Detection and Correction Punctuation Signs

6.1. Schemes of coordination

There are a différents schemes in texts :

Scheme (0) = [x1, x2, x3, ...x_i, ...x_{n-1}, x_n]

Scheme (1) = [x1, x2, x3, ...x_i, ...x_{n-1} AND x_n]

Scheme (2) = [x1 AND x2 AND x3 ... AND x_i ...x_{n-1} AND x_n]

Scheme (3) = [AND x1 AND x2 AND x3 ... AND x_i ...x_{n-1} AND x_n]

Scheme (4) = [x1, x2 , x3 , ...x3 , ...x_i , ...x_{n-1} , AND x_n]

Scheme (5) = [x1, AND x2, AND x3 , AND x_i ...x_{n-1} , AND x_n]

Scheme (6) = [AND x1 , AND x2, AND x3 , AND x_i ...x_{n-1} AND x_n]

6.2. Algorithm of Recognition the proximity of connectors

The algorithm of MAEGAARD&SPANG is cancels the **"coordination *information*"**, when the structure is : **[P1 C P2]** . This destruction (erase) of connectors C is used to obtain the "simples propositions". So, we lose the information of the presupposition and the temporal sentence. The problem is when the frequency of COMMA is high in the corpus of AFP, we must be distinguished the "comma of connectors" and the "free_comma".

Ç={AND; OR; Comma; Comma_OR; Comma_AND}

Phrase	Form	Frequency occurrences	Row of	Proximity
P_1	$Ç_i$	f_i	r_i	V_{i-p}/V_{i+p} (i,p∈[1,...,n])
.	. . .			.
.	$Ç_k$	f_k	r_k	V_{k-p}/V_{k+p} (k,p∈[1,...,n])
.	. . .			.
	$Ç_t$	f_t	r_t	V_{t-p}/V_{t+p} (t,p∈[1,...,n])
P_j				(j∈[1,...,n])
P_n				

The ranks are compared in the analysis of textual string, a reorganisation of the words by detecting the occurrences of commas in order to correct their value. To be able to determine the role of the comma, the following points have to be considered :

- The observation of the textual environment in terms of word classes, can it be helpful in the attempt to resolve ambiguities?

- The environment of the connectors is the neighbourhood (*PROXIMITY*) V_{i-p}/V_{i+p}, that is to say the words preceeding and following the comma;

- Is there a connector on the right side of the comma?

- The comma, does it connect predicates, noun phrases or propositions?

7. Logics and Statistical contribution in Computational Linguistics

In the framework of the analysis of written documents we have described the mechanism of substitution (correction) that allows to obtain (**LSI**) solutions, but another problem appears, the problem of distinguishing the value of the neutral comma compared to the other ones.

In a perspective of making our system operational we desire to compare the results obtained by working on the AFP corpus to other technical corpora with special attention to the indicator that is the *"frequency of coordinative commas"*.

The important volum in the scientific and technic information created a development of the automatic documentation. Indeed, the introduction of thousands of texts of any order in a automatic system contributed in giving to the written another dimension. But, among these writter documents, there are bodes (signs of punctuation). The coordination is ensured on the one hand by the natural connectors *"AND", "OR", " NOT", "BUT", "SO", "BECAUSE"*, and on the other hand by other processes of punctuation such as : *the comma, the semicolor, the dash, the brakets,...*

8. Conclusion

During the query formulation in natural language to obtain information in information system, the determination of the value of comma is important because it can replace the connectors {*"and", "or", "comma-and", "comma-or", "point-comma", "point-comma-and"*}. In thus, the development of tools to facilitate the access to Information System has to take account of this constraint in order to optimize the Man/Computer interaction because a query can contain one or several commas. That's why we make the claim that it is important to understand the functions of the comma in order to obtain results in the automatic analysis of documents. We think that statistical methods are optimal when accompanied by linguistics tools.

[1] **Croft (W.B), Lucia (T.J), Cohen (P.R);** *"Retrieving Documents by Plausible Inference : A Preliminary Study"*;In Y. Chiaramella (Ed.); ACM Conference on Research and Development in Information Retrieval; Grenoble; June 13-15; 1988; pp. 481-494.

[2] **Das-Gupta, P.** (1987), Boolean interpretation of conjunctions in document retrieval. *Journal of the American Society for Information Science, 38, 248-254.*

[3] **Dubois (D.), Prade (H.)**; *Théorie des possibilités;* Masson; Paris; 1985

[4] **Hintikka (J.);** *The Semantics of Questions and the Questions of Semantics*; North-Holland; Amsterdam; 1976.

[5] **Larouk (O);** *"Extraction de connaissances à partir de documents textuels : Traitement automatique de la coordination (connecteurs et ponctuation)"*. Thèse de Doctorat informatique, Université C. Bernard Lyon I, 1993. Thesis publisched in 1994.

[6] **Larouk (O.);***"An Evalution of the Textual Databases using Linguistic and Logics"*; in «First International Workshop on applications of Natural language to Data Bases» NLDB'95; *(Traitement des Langues naturelles et Bases de Données).* AFCET; Université de Versailles; June 28-29th 1995, pp. 150-158.

[7] **Larouk (O.), Bouché (R.);***"Apports des logiques et de la linguistique dans la conception d'interface de Bases de données textuelles"*; in **«Pluridisciplinarité dans les Sciences Cognitives»**; HERMÈS; Septembre 1993; Paris; pp. 142-160.

[8] **Le Guern (M)**; *"Un analyseur morpho-syntaxique pour l'indexation automatique"*; Le Français Moderne; tome LIX; n° 1; juin 1991; pp. 22-35.

[9] **Salton (G), Mc Gill (M.J);** *Introduction to modern information retrieval*; Mc Graw-Hill; 1983; New York.

[10] **Salton, G.** (1988); A simple blueprint for automatic boolean query processing; Information Processing&Management; 24; 269-280;

[11] **Spark Jones (K.)**; *"User models, discours models, and some others"*; in Y. Chiaramella (éd.); ACM/SIGIR: 11 th Conference on Research & Development in Information Retrieval; Grenoble; 1988; pp. 13-29.

[12] **Turner (R)**; *Logics for AI*; Dunod informatique; Paris; 1988.

[13] **Van Rijsbergen (C.J);** *"A non-classical logic for information retrieval"*; The Computer Journal; Vol. 29, n°6; 1986; pp. 482-485.

[14] **Zadeh (L)**; *"Fuzzy logic"*; IEEE computer; April; 1988.

[15] **Zadeh (L);** *"Test-score semantics as a basis for a computational approach to the representation of meaning"*; Literary Linguistic Computational; vol 1; 1986.

Adam Pawłowski
University of Wrocław

Language in the line *vs.* language in the mass. On the efficiency of sequential modelling in the analysis of rhythm.

Abstract

The subject of the present paper is the application of the **ARIMA** method of time-series analysis and of conventional statistical tests to the study of rhythm in text. The results obtained on the same set of samples with both investigative approaches are then compared. Sequential modelling by means of the ARIMA method turns out to be much more efficient than "mass" statistics.

Introduction

There are two complementary investigative approaches in quantitative linguistics: *the analysis of language in the mass* and *the analysis of language in the line*. The former considers linguistic units as statistically independent, regardless of their order in text. The latter is based on the assumption that their sequence is a relevant characteristic under investigation. **(...)** All these methods have been known for quite a long time outside linguistics but, surprisingly, the majority of quantitative studies of language have simply ignored the linearity of language, considering it as a mass phenomenon.[1]

Recently, more and more studies are devoted to the sequential analysis of language. In our opinion, one of the promising techniques to be applied in this field is the **ARIMA method of time-series modelling (...).** While the information theory takes into account relatively short sequences of linguistic units and the calculation of the entropy of higher orders encounters serious difficulties[2], the **ARIMA** method, based on the notion of syntagmatic time[3], allows the treatment of any series of linguistic data.

One of the crucial problems to be solved now is the efficiency of conventional, "mass statistics" and of sequential methods in the treatment of linguistic data. We can distinguish three cases here:

1. Only "mass statistics" can be applied.[4]

[1] Our remarks concern numerous works on model building as well as literary and stylistic applications of quantitative methods but they are not relevant for all the domains of language treatment (e.g. speech recognition).

[2] This question is discussed in the monograph of Rolf H. Hammerl and Jadwiga Sambor (1990, pp.370–404).

[3] "Nous allons appeler l'axe linéaire sous-jacent à la succession d'unités linguistiques temps syntagmatique (...) et nous allons le substituer au temps réel des événements observés. La pertinence de l'analogie entre ces deux notions est une condition essentielle de l'application de la méthode ARIMA au traitement du langage." (*Pawłowski 1997*, p.4)

[4] It's hard to imagine the contrary case where only sequential analysis could be applied, because a series of linguistic units (i.e. text or speech) can be always transformed into a set of units.

Examples of linguistic data which cannot be analysed as series are easily found if we remember that statistical populations in quantitative linguistics are basically divided into **text** and **system** populations (cf. *Hammerl, Sambor 1990*, pp.15–16). While the former can be represented either as sets (order of units is not relevant) or as series of elements, the latter simply ignores the notion of order and thus do not undergo sequential treatment.

2. Both approaches are possible and bring positive results.
The research carried out so far has proved, for instance, that the sequence of quantities of information conveyed by consecutive graphical words in French and in English text is not random but can be described as a *moving-average* stochastic process **MA(1)** (*Pawłowski 1997*, pp.96-106). Similar results have been obtained for the sequence of graphical word lengths in Italian (*Corduas 1995*). At the same time, reliable statistical models can be estimated in both cases and their parameters are likely to describe text samples in a satisfactory way.[5]

3. Both approaches are possible but only one of them gives satisfactory results.
It should be pointed out here that *linearity* is a fundamental characteristic of natural language and, at least in some cases, sequential modelling should turn out to be more efficient than other approaches.

Goal of the study

Our study will focus on the third case presented above. We intend to examine several samples of linguistic data by means of both conventional "mass" statistics and linear **ARIMA** modelling in order to compare the results obtained on the same set of data and decide which approach is more appropriate for research. Since this comparison will be based on the results of our previous study on the sequential structure of spoken Polish[6], detailed questions concerning the choice, the origin of the samples and the quantification will not be discussed here (cf. *Appendix 2*).

Results

The target of our analysis is spoken text in Polish transformed into a sequence of stressed and unstressed syllables, replaced with numbers 1 and 0 respectively. Samples represent four different types of versification (called *prosodic types*): **(...)** It is clear that these samples can be treated either as **series** or as **sets** of units and, at least in theory, both approaches are likely to reveal characteristics that would allow us to discriminate between them.

In the first case, text is transformed into a time-series submitted to sequential analysis in time- or frequency-domain. It can be shown (Tab.1) that each of the prosodic types mentioned above conveys a stochastic process which may serve as a

[5] Good results can be obtained with the logonormal distribution.
[6] Adam Pawłowski, Time-Series Analysis in Linguistics. Application of the ARIMA Method to Some Cases of Spoken Polish. [in:] *Linguistic Structures. To Honour Juhan Tuldava* (to be published).

"distinctive feature". The most striking differences occur between versified and non-versified texts, since in these cases quite different time-domain models of processes are estimated. However, insofar as more precise parameters are taken into account **(...)**, the **ARIMA** method discriminates between all the prosodic types analysed, including literary prose and rhetoric or political discourse **(...)**.

The table below (*ibid.*) includes the results of sequential analysis of different samples (cf. *Appendix 2*) by means of the **ARIMA** method. **(...)**

	Type of the model	Sample 1	Sample 2	Sample 3	Mean
Jan Brzechwa (simple verse)	**SARMA(1,1)$_4$**	57,8%	62,8%	68,8%	**63,1%**
Juliusz Słowacki (complex verse)	**SARMA(0,1)(1,1)$_{11}$**	45,0%	48,7%	50,6%	**48,1%**
John-Paul the 2nd (rhetoric discourse)	**MA(2)**	39,4%	43,1%	39,3%	**40,6%**
Igor Newerly (literary prose)	**MA(2)**	36,1%	31,3%	32,9%	**33,4%**

Tab. 1

But these analysed samples can be also considered as **sets** of elements where the order of accentuated / non-accentuated syllables is not relevant. The only characteristic we can determine in this case is the **percentage of accentuated syllables in text** (if 0's and 1's used for coding are treated as purely qualitative symbols, thus units of *nominal scale*) or the **mean of the sample** (if 0's and 1's used for coding are considered as numbers, thus units of *cardinal scale*).[7] Since it's difficult to decide which approach is more adequate, we will compare samples with regard to both characteristics and, consequently, statistical tests for the comparison of **fractions** as well as **means** will be used.

The statistic applied in the comparison of fractions of a given element in two samples (so called *coefficient of structure*) has a form:[8]

(1.) $$U = \frac{w_1 - w_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}}$$

where: **(...)**

We formulate the null hypothesis H_0 that both fractions are equal $w_1 = w_2$. Since the U-statistic has an asymptotically normal distribution N(0,1), the hypothesis H_0 is accepted on the level of 95% when the values of U belong to the critical interval $[-1{,}96;1{,}96]$. A positive result of the test (H_0 accepted) implies in this case that both samples come from the same statistical population and, consequently, are not statistically different.

[7] The problem of scale used in coding of syllables is discussed in the paper *Pawłowski 1997b*. Cf. also Note 15.
[8] Cf. *Hammerl, Sambor 1990*, pp.269–271, *Sobczyk 1996*, pp.166–167, *Greń 1987*, pp.419–424.

The test used for the comparison of two means has a form:[9]

(2.) $$U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where: **(...)**

As in the former case, we assume as the null hypothesis H_0 that both means are equal $\bar{x}_1 = \bar{x}_2$. Since the U-statistic has an asymptotically normal distribution N(0,1), the hypothesis is accepted on the level of 95% when the values of U belong to the critical interval [−1,96;1,96]. A positive result of the test (H_0 accepted) proves that there is no significant difference between the samples with regard to the mean.

In the table below, we present the result of one-to-one comparison of single samples. Numbers above the diagonal are the values of the U-statistic for the comparison of fractions in samples (0's and 1's are then considered as symbols on the *nominal scale*). Numbers below the diagonal are the values of the U-statistic for the comparison of means (0's and 1's are then considered as numbers). Shadowed values indicate statistically different samples. (cf. *Appendix I*)

	B1	B2	B3	S1	S2	S3	W1	W2	W3	N1	N2	N3
B1		0,393	-1,087	0,699	0,019	-0,152	0,638	0,261	0,070	0,599	0,682	0,643
B2	0,637		-1,384	0,167	-0,438	-0,588	0,118	-0,220	-0,383	0,084	0,161	0,150
B3	-1,653	-2,162		1,095	1,326	1,158	1,917	1,554	1,342	1,879	1,952	1,847
S1	1,115	0,273	2,072		-0,892	-1,118	-0,068	-0,563	-0,788	-0,117	-0,005	-0,011
S2	0,030	-0,717	1,964	-1,450		-0,227	0,805	0,319	0,067	0,755	0,858	0,783
S3	-0,240	-0,964	1,713	-1,810	-0,359		1,026	0,543	0,284	0,976	1,077	0,984
W1	1,025	0,194	2,585	-0,114	1,324	1,682		-0,484	-0,707	-0,048	0,061	0,051
W2	0,417	-0,363	2,326	-0,934	0,516	0,876	-0,812		-0,239	0,435	0,541	0,493
W3	0,112	-0,632	2,016	-1,302	0,109	0,458	-1,181	-0,394		0,660	0,760	0,698
N1	0,959	0,138	2,419	-0,194	1,233	1,587	-0,081	0,725	1,093		0,108	0,094
N2	1,091	0,264	2,93	-0,008	1,400	1,751	0,103	0,899	1,259	0,181		-0,006
N3	1,036	0,247	2,812	-0,018	1,289	1,614	0,085	0,823	1,162	0,158	-0,010	

Tab. 2

Notations:

B1, B2, B3 – samples of simple verse (children verses of Jan Brzechwa);
S1, S2, S3 – samples of complex verse (romantic poem of Juliusz Słowacki);
W1, W2, W3 – samples of rhetorical discourse (homily of John Paul the 2nd);
N1, N2, N3 – samples of literary prose (novel of Igor Newerly);

This result is confirmed by the comparison of summary samples for each prosodic type:

[9] Cf. *Hammerl, Sambor 1990*, pp.253–254, *Sobczyk 1996*, pp.155–158, *Greń 1987*, pp.419–424.

	B	S	W	N
B		0,775	1,002	1,517
S	1,213		0,322	1,015
W	1,582	0,527		0,688
N	2,395	1,670	1,149	

Tab. 3

When analysing the first of the tables above (Tab.2), we notice that except for one fragment of Jan Brzechwa's text (sample B3 – simple verse), there are no statistically significant differences between the samples. And even this peculiarity of Brzechwa's texts is not confirmed by both tests.

The summary chart (Tab.3) confirms this observation: **(...)**.[10]

As we can see, this result gives good grounds to claim that the distribution "in the mass" of stressed and unstressed syllables in Polish is constant and independent of the prosodic type of text.[11] This precious conclusion, however, remains in flagrant contradiction with the fact that the linear structure of analysed texts is **different.**

Conclusion

The result of our experiment proves serious deficiency of the conventional statistical tests in the analysis of rhythm. Not only do important characteristics of rhythm in single samples remain unknown, but also evidently different samples are found to represent the same statistical population. These difficulties do not arise if we apply sequential modelling by means of the **ARIMA** method. The linear structure of text is then revealed and different prosodic types can be distinguished.

The research carried out so far has proved that in the case of prose texts, good results are obtained with time-domain modelling. For instance, the sequence of stressed / unstressed syllables in Polish can be described by means of a simple **MA(2)** model. Versified texts are more regular and thus can be effectively treated both in time- and frequency-domain. In the first case, we obtain seasonal models, seasonality being equal to the length of verse. In the second case, spectral analysis detects the dominant frequencies in the spectrum of the series. Estimated model parameters, both in time- and in frequency-domain approach, can be then considered as "distinctive features" of each prosodic type. (cf. Tab.1 and *Pawłowski 1997b*).

[10] It's worth mentioning here that the test for the comparison fractions (values above the diagonal) is weaker than the test for the comparison of means (values below the diagonal). This is one more argument for the application of the binary ordinal scale in the research on rhythm in text.

[11] This statement seems to us plausible but still requires a thorough verification, as the number and diversity of treated samples are insufficient to establish more general laws of language structure.

ROUND TABLE DISCUSSION

Round table discussion
"Regularities in Natural Language Dynamics"

Chair of the round-table discussion:
Sheila Embleton (York University, Toronto, CANADA)

Introduction

Dear Participants of Qualico-97,

In order to activate discussion of some theoretical questions at our Helsinki meeting, Anatoliy Polikarpov has compiled, and we now suggest to you, a set of questions. In our opinion, they concern, some central and also "hot" points in recent developments in General Linguistics (and Quantitative Linguistics as an important part of it). Naturally, the range of questions reflects the spheres of interest of their authors. Nevertheless, raising them can facilitate the raising of others, which may have been overlooked or underestimated by the compilers. We are absolutely certain on one point - problems of cognition of language dynamics (both synchronic and diachronic) are noteworthy for our meeting, because regularities of processes and mechanisms of their mutual coordination within some real (i.e., developing) system is, from some point of view, a final goal of scientic activity. Meanwhile, a scientific approach to the cognition of processes and their coordination-correlation within some system mechanism urgently needs a means for estimating d e g r e e s (of parameters) and elaborating m e a s u r e s - natural scales of the processes within some system. That is why the significance of Quantitative Linguistics will increase in time for Linguistics as a whole.

We welcome preliminary reactions by Qualico-97 participants on the subjects raised. Please send your informal answers (specifying the number of the corresponding question) by e-mail to embleton@yorku.ca or polikarp@philol.msu.ru. You might also have suggestions for enlarging the questionnaire or the range of topics considered. We plan to prepare a summary of the responses received, to be presented at the conference.

We of course also welcome reactions and discussion at the conference itself -- this is after all the purpose of a roundtable discussion!

For a better understanding of some of the questions raised here we refer you to the various well-known works of G. Altmnann and R. Koehler, as well as to the two papers by A. Polikarpov submitted to the Conference.

Questionnaire

1) Are there mere "changes" or actual "evolution" (i.e., directed and coordinated reorganization) of the whole language system during its existence in time? If yes:

a) What is the reason for coordination of different parts and levels of a language in the process of historical development?

b) What are the factors (internal or external) in the evolutionary changes happening with various human languages?

2) Can you regard a mechanism of language communication, at the same time, as a mechanism of evolution?

If yes:
What specific features of it represent this phenomenon?

If no:
What are the real roots of language change in different typological directions?

3) Dealing with a mechanism of establishing equilibrium between contradicting "forces" ("factors", "requirements") in language existence, can you specify the situations of change in "weights" of some of these factors leading to establishment of a new state of equilibrium? Are there changes in some boundary conditions?

4) Do you agree that some major ("macroscopic") changes are the result of specific integration of some minor, "microscopic" changes, occuring with any micro-part of an object?

If so:
What are the micro-, meso- and macro- levels in language organization? What is the mechanism (and, possibly, specific conditions) for integrating micro-changes in some micro-dynamics? Are they not rooted in the very deep nature of a communicative act? Can you give examples for that?

5) What is the nature of a communicative (speech) act? What are its main features as a scene for communicative dynamics (exchange of ideas by communicants) and, at the same time, in evolutionary dynamics (loss and acquisition of some features by any language sign used in the communicative act)?

6) What other kinds of microprocesses in a sign history, besides the tendency of any sign meaning to abstractivization and generation of new (more abstract) meanings, can be found in the language realm?

7) Do you believe in some constant rate of core vocabulary change in time (only complicated by some stochastic and uncontrolled factors) as was claimed by M. Swadesh and his successors, or on the contrary, in some regular dependence of the rate on some features of a language's typology?

8) If there is inequality between different words in their ability to survive over time, what is the ground for the inequality in this ability and for the mechanism for words' falling out of a vocabulary?

9) Do you agree with a thesis on the irreversibility of some micro-processes in language existence? What are the consequences of this irreversibility?

Major Tendences in Micro- and Macro-Dynamics of Natural Language Lexical System

Anatoliy A. Polikarpov

Moscow Lomonosov State University
Faculty of Philology, Russian Language Dept.
Laboratory for General and Computational Lexicology and Lexicography

e-mail: polikarp@philol.msu.ru

1. There are some attempts to model Natural Language Lexical System change in time in quantitative aspect. They differ in language levels, subsystems attracting investigators' attention, in depth, in complexity of the subject, etc. The most important among them, in my opinion, were quantitative models elaborated by G.K.Zipf, M.Swadeshand M.V. Arapov. Zipf's model [1949] concerns some specific equilibrium of long-term language memory economy ("language", or "paradigmatic economy") of a typical speaker and operative apprehension economy ("speech", or "syntagmatic economy") of a typical hearer in speech activity. The main point is that equilibrium is a result of mutually opposed directions for "price" changes of each of them in the result of changes happened with a counterpart. But there was not put a question by G. K. Zipf on possible regular changes of "price" for different kinds of economy in different socio-communicative conditions resulting in changes of the whole optimum result. In my works [1976; 1979; 1986; 1987; 1993; 1994; 1995] these questions were put on the ground of System Linguistics approach [Melnikov, 1978; 1988]. The main point was that in the sutuation of noticeable spread of some language on a wider community (on native speakers of other languages) the "price" for "paradigmatic economy" for an average communicant should increase. It is a result, first, of impossibility for nonnative speakers to master their new language skills immediately. It, further, leads to often omission of some specific categories of signs from native speakers speech, i.e., to narrowing of the commonly used core of language signs and constructions.

This stage of the adaptive process could be called "speech adaptation". Stabilizing of the situation in 2-3 generations leads, eventually, to fixing of preserved sign units and the following reconstruction of relations between them in the form of redistribution of the overall set of semantic functions among preserved sign elements, and to beginning of adaptation of preserved elements to newly acquired bunches of functions (usually much more volumous for each of remaining signs, even in the absense of some most specific, idiosyncratic language functions completely thrown away by the new communicative practiceas unaffordable now for a new "average speaker").

This stage of the adaptive process could be called "language adaptation". It is realized for language elements (morphemes, words and phrases) in some stochastically formed order. Two main selective criteria are present in the process.

The first, affixes, grammar words and grammar expressions are difficult for use, due to specific grammar idiomaticity present in their meanings. The more idiomatic meanings, the greater difficulty of acquisition of them for nonnative speakers and less chances for preserving them is present in a system.

The second, rare words covering some specific (or peripheral) areas of the whole semantic space of a community are difficult for use and acquisition. The rarer a word, the greater is a difficulty for an average speaker to obtain knowledge of its meanings.

Disappearing of the situation of intensive language spread (as a result of growing homogeneity, i.e., levelling individual language knowledges during the process of mutual - "plus" and "minus" - teaching of elements and patterns present in communication) inescapably leads to arising of the reverse process - of syntheticization of a language system. This is naturally determined by the back change in the "price" of each mentioned kinds of "economy" in favour of greater significance in new conditions of the "syntagmatic economy". Newly acquired equilibrium of "prices", where the whole cost resulting from summing up of two types of economy is an optimizing parametre, limits continuation of changes in any direction.

This view now is elaborated further within the model of the word life cycles [Polikarpov, 1988; 1993; 1994; 1995; 1997]. See some additional components for it in the abstract of my paper "Semasiological and Word-formational Processes in Natural Language Lexical Evolution" also submitted for the Qualico97.

2. Some other trend in quantitative modelling of language historical development is present by works of M.Swadesh and his successors. Glottochronological model by M.Swadesh up to now is the most popular view of some general regularities of lexicon survival in time. It, really, had opened a new page in formal description of some global tendencies of language (here - lexical) development. But it appeared to be weak in explanation of different rates of a core vocabulary decay for different languages and that of a vocabulary of the same language in different historical periods. Moreover, there was not even put a question about some possible regularities for the decay process of different parts of an overall vocabulary of some language, or, at least, for a core vocabulary (100-200 words). Namely, there was not even drop a hint on some possible dependence of words' historical fate (more or less safe existence in time) on their own grammatic, semantic and other system features. Basing on the ideas of the word life cycle model the last problem was experimentally studied by M. Kapitan (see, for instanse, his paper in JQL, 1994, v.1, N3 [Kapitan, 1994]) using data from the history of Romance languages - from classical Latin to modern Italian, Spanish, Portugese, French, Rumanian and some others. Analysis of Slavic, Germanic and other languages data on this point are also in progress at the Laboratory for General and Computational Lexicology and Lexicography of Moscow University.

At last, system regularities for the process of replacement of words leaving the vocabulary by new (entering) words and for their further existence-development was not considered in Swadesh's model at all. So, the whole problem of the system process of vocabulary renovation and evolution was not even put.

On the whole, Swadesh's model was highly restricted onthologically, made a stress only on the phenomenology of words' falling out, not explaining it, not even hinting on possible typological causes and mechanizms of different languages' words falling out rate. That is because the process of "decay" was not integrated in some more complex onthological (communicative) picture of language existence (including lexicon renovation process varying in its degree for languages of different typology and languages existing in different communicative conditions).

3. A specific model for the integration of various factors of language existence-evolution, called a model of the word life cycle, takes as its initial postulate presence of some specific ability in any language sign (say, a word) - so called

associative-semantic potential. The potential is manifested by the more explicitely presented ability of any meaning to enter into assosiative links with senses and other meanings which forms the basis for the principal hinting-quessing mechanizm (sense - meaning - sense) acting in Natural Language communication. Associative-semantic, or, simply, semantic potential of a word is determined and can be measured by the degree of its first meaning's concreteness, "dencity" of semantic components contained in it. During time life, being used in communication, a word gradually spends the potential. Spending of it manifests itself basically in two main processes.

First, in change of any word meaning during its communicative life-use (beginning from the first meaning) in the direction of its growing abstractness, i.e., loss of semantic components while inescapable broadening of its sense scope during each communicative act. It means that in any act of communication different speakers are, naturally, trying to apply meanings to slightly different areas than they previously used to. It leads, correspondingly to the neutralizing and, eventually, omitting of those components of any meaning which become unrelated in time to any component of senses covered by it.

Second, spending of the potential manifests itself in attraction-acquisition by any meaning of new meanings (in the process of fixing some most useful in this respect associative links of a meaning with senses and converting some of them into new meanings). Spending in this case means making of some components of some maternal meaning already busy by some associative links and therefore unable (or less able) for the same converting other senses into meanings.

Besides, it should be taken into account, that each of new meanings should be, on the average, more abstract than each parental one. It is predetermined by the evolutionary preference for a meaning (as compared to other possible ways) to acquire not any meanings, but, better, those meanings which are relatively more abstract and the refore - more stable than any other among possible senses-candidates for a role of a new meaning. Greater stability of more abstract meanings can be explained by their ability to cover a broader sense sphere and therefore - by their lesser dependence on changes in any local extralinguistic (sense) sphere of the meaning as compared to any more concrete meaning.

Abstractivization of any meaning in its history and getting by a word of more and more abstract meanings in its history are two kinds of processes determining basic micro-dynamics of any word (and, moreover, dynamics of any sign of other language levels in their own history). Abstractivization processes, as the strongest ones, predetermine any other microprocesses in any word life cycle. This, naturally, presupposes arising of all basic lexical system regularities, i.e., regularities in macro-dynamics of the whole vocabulary of a language.

Individual variations of the semantic potential among words concern the degree of activity of any word initial meaning (its productivity in giving birth to new meanings) and stability (its ability to resist to unfavorable factors, to exist some certain time not falling out from language). This determines level of activity and stability of each next word meaning (because of the mentioned above regular dependence of the ability of any next meaning on the ability of its predecessor) and the life cycle of the word on the whole - involving many other language system parameters: semasiological (synonymy, antonymy, homonyny, etc.), phraseological, morphemic, word-formational, flectional, phonological, frequency of use, length, etc.

Individual variations of the semantic potential for the whole set of words entering

language at some certain period of time are not chaotic, they should follow some distribution pattern. Most of words should be extremely unstable and inactive. The greater the level of word activity or stability, the smaller proportion of words entering some moment into a language posesses this level of them. The rarest among others should be words extremely active and at the same time extremely stable.

4. Important step in quantitative modelling of language evolution consists in assuming the idea of irreversibility of mentioned parameters change for any word in its history. It determines any vocabulary (and language on the whole) inescapable renewal. In the case of stable communicative conditions it leads mainly to the replacement, proportional renewal of elements without noticeable change of structural features of a vocabulary, and that of a language on the whole. But according to changes in some typologically relevant comunicative conditions (e.g., arising of a mentioned above significant spread of some language on some nonnative speakers of it or arising after some period of intencive language "mixture" of an opposite situation of stable community functioning with the absense of the noticeable ethnic mixture for some relatively long period of time and corresponding growth of the degree of language homogeneity of it) this eventually leads to some significant change of its typological shape, e.g., analytic or back - synthetic - restructuring. In the case of analytical development and corresponding shrinking of lexical vocabulary communicants redistribute the whole bulk of lexical functions between remaining lexical items which leads to the increase of the average functional load for each of them - increase of the number of meanings and frequency of use. Naturally, it further leads to increase of the average speed of a word life cycle, faster wearing out, on the average, of each of the units of the whole lexical system and, correspondingly, to faster (but with specific coefficients) wearing out of units of the root and affixal systems of language.

5. Someone can observe in objective language reality different rates of vocabulary replacement in different communicative conditions according not only to mentioned above factors, but according also to socially determined changes in the size of the sense sphere covered by a language, shrinking or rising of some language use in the same spheres (e.g., in the situations of becoming dominant or, on the contrary, oppressed language in a multiligual society), shrinking or rise of the community size, etc., and, of course, according to various combinations of all of these factors.

All this, seemingly, is in a clear contradiction with claims of M. Swadesh and some of his successors on some constant, universal norm of changes for any vocabulary at any time. Real linguistic evolutionary mechanizms are communicative in their nature and should be studied beginning from the microlevel of their organization. Only this, with the combination of the information on relevant boundary conditions of a community existence, can give an opportunity to approach closer, than before, to the understanding of real mechanizms of language life and evolution, to understanding language system tendences and laws.

REFERENCES

ARAPOV M.V., KHERTS M.M. [1974].
Mathematical Methods in Historical Linguistics. - M.: Nauka, 1974.

BREITER M.A. [1994].
Length of Chinese Words in Relation to their other Parameters // Journal of Quantitative Linguistics. V.1, N3, 1994.

BREITER M.A., POLIKARPOV A.A. [1997].
Polysemy and Frequency of aWord in Chinese: Experimental Study of System Dependences // IV International Conference on Languages of the Far East, South-East Asia, and Western Africa, September 17-20 1997, Institute of Asian and African Countries, Moscow Lomonosov State University. M., 1997.

KAPITAN M.E. [1994].
Influence of Various System Features of Romance Words on their Survival // Journal of Quantitative Linguistics. V.1, N3, 1994.

MELNIKOV G.P. [1978].
Systemology and Cybernetic Problems in Linguistics (in Russian). M.: Sovetskoe Radio, 1978.

MELNIKOV G.P. [1988].
Systemology and Cybernetic Problems in Linguistics. -L.- Sidney: Gordon and Breach, 1988.

POLIKARPOV A.A. [1976].
Factors and Regularities of the Analiticity Development in Language. Ph.D. dissertation. Moscow university. - M., 1976.

POLIKARPOV A.A. [1979].
Elements of the Theoretical Sociolinguistics (in Russian).- M.: Moscow University Press, 1979.

POLIKARPOV A.A. [1986].
On the Notion of "Communicative Situation" // Methodological Problems of the Social Linguistics (in Russian). - M.: Moscow University Press, 1986.

POLIKARPOV A.A. [1987].
Polisemiya: Sistemno-Kvantitativnyye Aspekty. (Polysemy: Systemic-Quantitative Aspects) // Quantitative Linguistics and Automatic Text Analysis (in Russian). - Tartu: Tartu University Press, 1987.

POLIKARPOV A.A. [1988].
K Teorii Zhiznennogo Tsikla Leksicheskikh Yedinits (Towards the Theory of Life Cycle of Lexical Units) // Applied Linguistics and Automatic Text Analysis. Papers from the All-Union Conference held 28.01-30.01.1988 at Tartu University).- Tartu: Tartu University Press, 1988.

POLIKARPOV A.A. [1993].
A Model of the Word Life Cycle // Contributions to Quantitative Linguistics / Ed. by R. Koehler, B.B. Rieger. - Dordrecht: Kluwer, 1993.

POLIKARPOV A.A. [1994].
Zakonomernosti zhiznennogo tsikla slova i evolutsija jazyka. Statja 1. Modelirovanije osnovnykh sistemnykh sootnoshenij (The Regularities of Word Life Cycle and Language Evolution. Article 1. The Modelling of the Main System Correlations) // Russkij Filologicheskij Vestnik (Russian Phylological Bulletin), N 1, 1994. - Moscow, 1994.

POLIKARPOV A.A. [1995].
Zakonomernosti zhiznennogo tsikla slova i evolutsija jazyka. Statja 2. Teorija i eksperiment (The Regularities of Word Life Cycle and Language Evolution. Article 2. Theory and Experiment) // Russkij Filologicheskij Vestnik (Russian Philological Bulletin), N 1, 1995. - Moscow, 1995.

POLIKARPOV A.A. [1997].
Lexical Sybsystem of Natural Language System: Theoretical and Experimental Aspects of its Coming-to-Be and Evolutionary Study (in Russiana manuscript). - Moscow, 1997.

POLIKARPOV A.A., KURLOV V.Ya., [1994].
Stylistics, Semantics, Grammar: Experience of System Correlations Analysis (On the Basis of Data from the Explanatory Dictionary) // Voprosy Jazykoznanija (Journal "Linguistic Problems"). N 1, 1994 (in Russian).

ZIPF G.K. [1949].
Human Behavior and the Principle of the Least Effort. - Boston (Mass.): Addison-Wesly, 1949.

FORMS

Type-based and Token-based Learning of Kanji Morphemes[1]

Kyo KAGEURA
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello St. Sheffield S1 4DP, UK
E-Mail: k.kageura@dcs.shef.ac.uk (until 25th December 1996)
National Center for Science Information Systems,
3–29–1 Otsuka, Bunkyo-ku, Tokyo, 112, Japan
E-Mail: kyo@rd.nacsis.ac.jp (from 1st January 1997)

Summary

We have been developing methods of kanji morpheme analysis for the empirical modelling of terminology. In this paper we discuss the performance of kanji morpheme extraction and kanji sequence decomposition, both based on the same bigram statistics, focusing on the effect of type-based and token-based trainings. The experiment shows that type-based training gives consistently better performance, which has both practical and theoretical importance.

Topical Paper : 2. Application of Quantitative Methods to Natural Language Processing

1 Introduction

We are currently carrying out a research project AMANITA (Automatic/ing Morpheme Accumulator/ion for NIhongo Terminological Analysis), under which we are establishing a theory of distribution of terminological elements in Japanese technical texts as well as developing methods and tools for analysing basic morphological units in the texts of a given domain.

As part of this, we are developing a quantitative method for analysing kanji (Chinese character) morphemes, which constitute the major part of Japanese terminology. There are two basic tasks: syntagmatically decomposing kanji sequences into proper morphemes (note that Japanese does not have orthographic boundaries of word or morpheme), and weighting the morphemes according to their 'importance' in the domain.

We have been examining the applicability of simple character based bigram statistics for these tasks[2]. After examining several measures such as X^2, likelihood ratio test, Yule's coefficient of colligation Y, and mutual information, we found that the likelihood ratio test performs quite well for weighting base morphemes according to their 'importance' in the domain, as well as for decomposing kanji sequences (Kageura 1996a).

[1]The author would like to thank Mr Rob Collier of the Department of Computer Science, University of Sheffield, for giving detailed comments on the draft.

[2]With very few exceptions such as some proper names and coordinations, kanji sequences consist of combinations of base morphemes (morphemes consisting of two kanji characters) and affixes (consisting of one kanji character). Thus character based bigram statistics are expected to be applied to kanji morpheme analyses straightforwardly.

In Kageura (1996a), the evaluations of the different bigram measures were carried out on the basis of token-based learning[3]. This is in accordance with the fact that most current statistical or quantitative NLP systems adopt token-based learning[4]. However, if the statistical measure can treat rare events properly, it is expected to obtain even better performances by type-based learning for the analyses of terminological elements, because formations of complex terms follow their own motivated rules independent of their actual use in the texts (Sager 1991). As the likelihood ratio test is sensitive to rare events (Dunning 1993; Sprent 1993), it was expected that it will give a reasonable performance on the basis of type-based training. In this paper we report the results of experiments for kanji morpheme analyses by type-based and token-based learning.

2 Likelihood Ratio Test for Bigram Collocations

The likelihood ratio test for bigram collocations, adopted as the basic measure for the kanji morpheme analyses, is defined using a two by two contingency table as shown below.

		Column variable = Second word		
		Category 1 (w_2)	Category 2 ($\overline{w_2}$)	Total
Row variable = First word	Category 1 (w_1)	$f_{11} = f(w_1w_2)$	$f_{12} = f(w_1\overline{w_2})$	$f_{1.} = f(w_1)$
	Category 2 ($\overline{w_1}$)	$f_{21} = f(\overline{w_1}w_2)$	$f_{22} = f(\overline{w_1}\overline{w_2})$	$f_{2.} = f(\overline{w_1})$
	Total	$f_{.1} = f(w_2)$	$f_{.2} = f(\overline{w_2})$	$f_{..} = \sum f(w_i)$

In the table, the row variable is the word in the first position of a bigram. The first category of the row variable consists of the word in focus w_1, while the second category consists of the word other than w_1 (denoted by $\overline{w_1}$). The column variable is the word in the second position, consisting of two categories, w_2 and $\overline{w_2}$. f indicates the observed frequency. Thus the top left cell f_{11}, for instance, whose value is indicated by $f(w_1w_2)$, shows the observed frequency of the collocation w_1w_2.

Using the notation in the above table, the likelihood ratio is defined as follows:

$$-2\log\lambda = 2[\sum_c \log L(f_{1c}/f_{.c}, f_{1c}, f_{.c}) - \sum_c \log L(f_{1.}/f_{..}, f_{1c}, f_{.c})]$$

where

$$\log L(p, n, k) = k\log p + (n-k)\log(1-p)$$

Because the value of the likelihood ratio test does not indicate whether the association between two variables is positive or negative, we used the sign of Yule's Y to distinguish positive and negative associations. Although the value of the likelihood ratio test becomes larger as the total ($f_{..}$) becomes larger, given the same ratio among cells, we did not 'correct' it because, in our tasks, bigram collocations are compared within the same corpus.

[3]In token-based learning, if a sequence type 'ABCD' (each capital letter indicates a kanji) appears ten times and 'ABD' twice in the corpus, for instance, 'AB' is counted 12 times, 'BC', 'CD' 10 times, and 'BD' twice, in calculating bigram statistics. In type-based learning, on the other hand, 'AB' is counted twice, and 'BC', 'CD' and 'BD' once.

[4]Note for some units, such as sentences, the distinction of token-based and type-based learning is unapplicable. However, even when applicable, type-based learning is not used, e.g. Hidden Markov Model based method for decomposing kanji sequences in Takeda & Fujisaki (19[illegible]).

3 Data for Experiments

In the following experiments, we used three corpora of different domains, forestry, artificial intelligence, and information processing, which were randomly chosen from the 'Database of Japanese Academic Conference Abstracts', a database serviced by the National Center for Science Information Systems, Japan. Kanji sequences extracted from each of these three corpora were used for the experiments.

The numbers of kanji sequences by length for each domain are as follows:

No. of Characters	Forestry		Art. Int.		Inf. Proc.	
	Type	Token	Type	Token	Type	Token
1	579	9038	556	13030	1190	196032
2	2501	14250	1974	27394	6866	389042
3	2208	4574	1824	6286	12383	88815
4	2565	4094	2844	6303	22653	83150
5	1057	1427	995	1423	10721	21499
6	564	692	661	977	8232	14150
7	252	289	231	332	3012	4509
8	111	124	111	137	1568	2160
9 ≤	127	152	75	87	1128	1571
Total	9964	34640	9271	55969	67753	800928

The numbers of different bigram collocations and of different base morphemes, i.e. morphemes consisting of two kanji characters, are as follows:

	Forestry	Art. Int.	Inf. Proc.
No. of Different Bigrams	10534	8393	42329
No. of Different Base Morphemes	3992	1819	9442

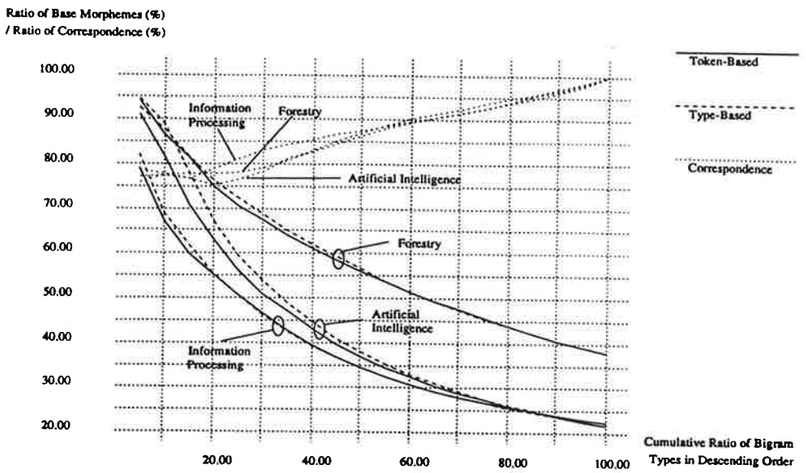
4 Weighting Base Kanji Morphemes

Weighting base kanji morphemes, consisting of two kanji characters, is similar to collocation extraction (Church and Hanks 1990; Smadja 1993; Dunning 1993). To evaluate the performance of the likelihood ratio test as the measure of weighting base morphemes, we calculated the value of bigrams for each domain by both type and token information, ordered them according to their values, and evaluated the percentage of linguistically valid bigrams (base morphemes) of the top 5%, 10%, etc[5].

Figure 1 shows the results of token-based and type-based learning for three domains. The horizontal axis indicates the cumulative percentages of the bigrams from the top according to the order of their weights, and the vertical axis shows the ratio of base morphemes among the bigrams. The figure also shows the ratio of corresponding bigrams between type-based and token-based weighting for each domain.

[5]The principal purpose of weighting bigrams in AMANITA is to evaluate their importance as terminological elements, thus qualitative evaluation from a terminological point of view is needed (for part of this see Kageura 1996b), we give here only a quantitative evaluation.

Fig. 1 Ratio of base morphemes and ratio of correspondence among bigrams from the top



Excluding the top 5% in the field of forestry, type-based learning of bigram weights performs consistently better than token-based learning, in the sense that the former gives a higher ratio of base morphemes among more heavily weighted bigrams. The statistical significance at the 95% level is observed around top 10% to 20% in artificial intelligence.

The ratio of corresponding bigrams between type-based and token-based learning is between 75% and 80% among the top 5% of bigrams for all three domains, and 85% to 90% of correspondences among the top 50%[6]. Although the qualitative evaluations of type-based and token-based weighting is yet to be pursued fully, bigrams scoring well in type-based calculation seem at least as good as those in token-based calculation.

5 Decomposition of Kanji Sequences

The method for decomposing kanji sequences is based on a very simple idea: if proper weights are given to all the kanji character bigrams, they should be used for measuring the strength of neighbouring bigram collocations within their actual occurrences in kanji sequences.

Based on this simple idea, the following algorithm is used to decompose kanji sequences:

```
1. Calculate the score for each bigram using the likelihood ratio test
2. Using the scores, apply the following procedure:
        decompose_string (string) {
            if (length of string <= 2) {
```

[6]Within highly ranked bigrams, the correspondence is much higher than correspondences between different measures. For instance, the average correspondence in the three domains between the likelihood ratio and the other measures based on tokens are: 56.75%, 78.91% and 96.34% for the top 5%, 20% and 50% respectively, between the likelihood ratio test and X^2, 19.67%, 60.95% and 90.36% for the top 5%, 20% and 50%, between the likelihood ratio test and Yule's Y, and 16.51%, 59.53% and 90.22% for the top 5%, 20% and 50%, between the likelihood ratio test and mutual information.

```
            return string;
        } else {
            divide string into head and tail,
                at the point where score is minimum;
            decompose_string (head);
            decompose_string (tail);
}
```

The result of the decomposition, applied to kanji sequences of length three or more characters, are as follows[7]:

	No. of	Type-Based Learning		Token-Based Learning	
	Sequences	Success	Percentage	Success	Percentage
Forestry	11352	9580	84.39	9604	84.60
Art. Int	15545	14852	95.54	14696	94.54
Inf. Proc	215854	205647	95.27	204756	94.86

In artificial intelligence and information processing, type-based learning performs significantly better than token-based learning. Only in forestry token-based learning performs better, but the performance difference is not statistically significant. The overall average accuracy for the data used here is 94.78% by type-base training, and 94.36% by token-based training.

If we ignore the relevance of weighting morphemes in each domain and purely focus on the performance of decomposition, it is possible to use the likelihood ratio values calculated by the mixed data. The following figure shows the result of decomposition based on the bigram values calculated for all the three domains together:

	No. of	Type-Based Learning		Token-Based Learning	
	Sequences	Success	Percentage	Success	Percentage
Forestry	11352	9761	85.98	9505	83.73
Art. Int	15545	14883	95.74	14704	94.59
Inf. Proc	215854	206425	95.64	204868	94.91

In this case, type-based learning outperforms token-based learning in all three domains. The average accuracy by type-based learning is 95.19%, while by token-based learning it is 94.37%. In any case, type-based learning performs better than token-based learning in decomposition.

This result is comparable to the Hidden Markov Model based decomposition developed by Takeda & Fujisaki (1987), the best known performance for kanji sequence decomposition so far, but much less training data. Their method gives an average performance of 95% with more than one million characters training data[8], and 97% with heuristics. With a few general post-processing heuristics applied to the result of type-based decomposition, our method gives an average of 96.89% accuracy for domain dependent training, and 97.00% accuracy for mixed domain training.

[7]Figures are based on running sequences.

[8]Whether their method works properly by type-based learning is not reported

6 Conclusion

The results of the above experiments show that, by choosing an appropriate statistical measure, we can obtain significantly better performance using type-based learning for analysing Japanese kanji morphemes. This has some practical importance for language processing applications. Firstly, as the size of training data is smaller in type-based learning, the processing efficiency improves. Secondly, it is possible to analyse the data which does not carry token information, such as entries of dictionaries, etc., which is very important for the AMANITA project, because there are many machine-readable lists of technical terms, which constitute very useful background resources for quantitative terminological research if they can be properly analysed.

So far we have been discussing the difference between type-based and token-based learning from the point of view of performance in morphological analyses. However, although proper investigation is required, the result can be interpreted from a linguistic point of view, i.e. of systematicity of structure of simple and complex lexical items *per se*, as distinct from their actual usage in the discourse. We are currently examining this aspect as well, in relation with the possibility of estimating the performance of morpheme analyses.

References

Church, K. W. and Hanks, P. (1990) "Word Association Norms, Mutual Information and Lexicography," *Computational Linguistics.* 16(1) p. 22–29.

Dunning, T. (1993) "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics.* 19(1) p. 61–74.

Kageura, K. (1996a) "Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences," Internal Memo, Department of Computer Science, University of Sheffield.

Kageura, K. (1996b) "Statistical Characterisations of Terminological Elements in Japanese Technical Abstracts with Reference to the Field of Artificial Intelligence," Internal Memo, Department of Computer Science, University of Sheffield.

Sager, J. C. (1990) *A Practical Course in Terminology Processing.* Amsterdam, John Benjamins.

Smadja, F. (1993) "Retrieving Collocations from Text: Xtract," *Computational Linguistics.* 19(1) p. 143–177.

Sprent, P. (1993) *Applied Nonparametric Statistical Methods* (2nd ed). Chapman and Hall, London.

Takeda, K. and Fujisaki, T. (1987) "Automatic Decomposition of Kanji Compound Words Using Stochastic Estimation," *Transactions of Information Processing Society of Japan*, 28.9:952–961.

Relating Word Length to Morphemic Structure: A Morphologically Motivated Class of Discrete Probability Distributions

– Topical Paper (Short Paper) submitted for QUALICO-97 –
(topic area: word length / probability theory)

PETER MEYER
Mittelstraße 2
D-37077 Göttingen
Germany

Georg-August-Universität Göttingen

Abstract. In this paper, a simple mathematical model for the distribution of word length in texts is proposed. Assuming that words may be analyzed into morphemes, such that both the number of syllables per morpheme and the number of morphemes per word follow the Poisson distribution, one obtains a class of probability distributions, some of which have turned out as useful models for texts written in Eskimo, an agglutinative, morpheme-based language.

1. Word length and morphemic structure: basic assumptions

Most previous attempts at modeling word length distribution in texts did not consider any specifically morphological properties of the languages in question. Hence, it is natural to ask whether it is possible to *supplement* standard, e.g. synergetic, approaches with some account of the morphology involved. Thus, word length in terms of syllable number in a strongly *agglutinative* language might be viewed as depending (a) on the distribution of *morpheme* length in terms of *syllable number* – henceforth *syllabic distribution* – and (b) on the distribution of *word* length in terms of *morpheme number* – henceforth *morphemic distribution*. As for the two 'composing' distributions, the research work done so far suggests assuming some member of the Poisson family. Indeed, I shall simply posit the simple, possibly displaced, one-parameter Poisson distribution as a model for both the morphemic and the syllabic distribution. These reflections result in two basic assumptions:

1. The morphemic distribution (i.e., distribution of word length as expressed in number of *morphemes*) is a simple, c-displaced Poisson distribution with parameter b, where $b > 0$ and $c = 0, 1, 2, \ldots$.
2. The syllabic distribution (i.e., distribution of morpheme length as expressed in number of *syllables*) is a simple, d-displaced Poisson distribution with parameter m, where $m > 0$ and $d = 0, 1, 2, \ldots$.

The average number of morphemes per word is easily seen to be $b+c$, whereas mean syllable number per morpheme amounts to $m+d$. Linguistically, values bigger than 2 or 3 for the displacement variables c and d are hardly plausibly, however. Mutual independence of all pertinent random variables is assumed for all deductions that follow.

It should be emphasized at the very outset that the above assumptions fall short of giving a sufficient characterization of the factors that must be expected to exert influence on word length distributions actually observed in texts.

2. The resulting class of probability distributions

The above assumptions suffice to generate a class of probability distributions that model word length in terms of syllable number. In this preliminary paper, I shall only give the bare

outlines of the mathematical reasoning. Let P(b, c, m, d, l) denote the resulting word length distribution, where b, c, m, d are defined as above and l is the number of syllables (the word length), $l \geq cd$. Let $\pi_d(a,x) = e^{-a}\dfrac{a^{x-d}}{(x-d)!}$ denote the d-displaced simple Poisson distribution with expectancy value $a+d$, where $x = d, d+1, \ldots$ The probability generating function (pgf) of P(b, c, m, d, l) can be obtained by functional composition of the pgf's of the composing distributions. Since the pgf of $\pi_d(a,x)$ is $G_{a,d}(s) = s^d \cdot e^{a(s-1)}$, the pgf $G_P(s)$ of P(b, c, m, d, l) is

$$(1) \qquad G_P(s) = G_{b,c}\left(G_{m,d}(s)\right) = s^{cd} \cdot e^{cm(s-1)+b\left(s^d e^{m(s-1)}-1\right)}.$$

To find a way to calculate the values of P(b, c, m, d, l) directly, we observe that

$$(2) \qquad \mathrm{P}(b, c, m, d, l) = \sum_{i=c}^{\omega}\left(\pi_c(b,i)\cdot \sum_{\substack{\langle n_1,\ldots,n_i\rangle \\ n_x=d,d+1,\ldots \\ \sum n_x = l}} \prod_x \pi_d(m,n_x)\right)$$

Here, the outer sum adds up probabilities for all possible numbers i of morphemes a word of l syllables can possibly consist of. The meta-symbol ω is taken here to stand normally for the integer part of l/d; if d=0, however, it has to be replaced by the infinity symbol ∞. The inner sum, henceforth abbreviated as φ(m, d, l, i), sums up probabilities for all possible „morpheme configurations", i.e., possible partitions of l syllables among i ordered morphemes with a minimum syllabic length of d, where m is the parameter of the syllabic distribution as defined above.

φ(m, d, l, i) can be represented by employing its pgf; we get

$$(3) \qquad \varphi(m, d, l, i) = \frac{d^l}{ds^l}\left(s^d \cdot e^{m(s-1)}\right)^i\Big|_{s=0} \cdot \frac{1}{l!}$$

Insertion of this into (1) gives, after some tedious calculation, the following formula, which defines a class of two-parameter (b, m) discrete probability distributions generated by inserting arbitrary integer values for c and d:

$$(4) \qquad \mathrm{P}(b, c, m, d, l) = \frac{e^{-b}}{l!}\sum_{i=c}^{\omega}\frac{1}{(i-c)!}\binom{l}{di}\cdot(di)!\cdot b^{i-c}\cdot(im)^{l-di}\cdot e^{-im}$$

For zero displacement values (c, d = 0), (4) becomes identical to the well-known Neyman distribution type A, viz.

$$(5) \qquad \mathrm{P}(b, 0, m, 0, l) = \frac{e^{-b}\cdot m^l}{l!}\cdot\sum_{i=0}^{\infty}\frac{i^l\cdot\left(be^{-m}\right)^i}{i!}$$

With both displacements set to 1, formula (4) turns into

(6) $$P(b, 1, m, 1, l) = \frac{e^{-b}}{l!}\sum_{i=1}^{l}\binom{l}{i}\cdot i \cdot b^{i-1}\cdot (im)^{l-i}\cdot e^{-im},$$

which does not belong to the Neyman family of distributions.

3. Empirical relevance of the distributions

Hitherto, the two members (5) and (6) of the class have been tested for empirical adequacy, if only for a small corpus of twenty traditional oral Eskimo narratives.The language of these texts is the Nunavik dialect of the Inuktitut branch of the Eskimo language family. A highly agglutinative language with complex morphemic structure, Eskimo seems to be a particular qualified candidate for the basic assumptions proposed above. Out of the texts checked, 18 could indeed be fitted to the (usually one-displaced) distributions in (5) and (6) ($P(X^2) \geq 0.01$); of these, at least 13 may even be said to fit well ($P(X^2) \geq 0.05$). It is likely that improvements in the fitting techniques employed might lead to considerably better results. Thus, my preliminary analysis did not make use of pooling classes. – It must be noted that all checked texts can be fitted to the Hyperpoisson distribution and that (5) and (6) can often be used as a good approximation for this distribution.

The final version of the present paper will contain fitting data for more Eskimo texts as well as for some further text corpora in other languages.

4. Further perspectives

I shall merely note some important points that still need consideration with respect to the probability distribution class defined in (4). I expect to include at least some of them in the final version of this paper.

- As said before, the empirical usefulness of the distribution class must be demonstrated on the basis of a more extended text corpus including materials from different languages. It would be interesting to see whether, e.g., agglutinative languages may, in general, be fitted more readily to these distributions than, say, more isolating ones. In view of the utterly reductive character of the proposed model, however, this does not seem to be very plausible.
- The empirical adequacy of positing a simple Poisson distribution for both morphemic and syllabic distribution must be tested against reasonable alternatives, particularly the Borel distribution, which has already proven to be a valuable instrument in word length modeling.
- In the proposed model, the two parameters receive a direct linguistic interpretation; this is perhaps its most salient feature. To take an example: In (6), b is the average morpheme number per word minus one, and m is the average syllable number per morpheme minus one. It should be tested whether the values obtained through the fitting process do indeed correspond to observable average syllabic or morphemic lengths, in spite of severe methological objections that such a hypothesis has to face.
- It would be important to include some details on the fitting mathematics that is adequate for the distributions in question.
- A very difficult problem is the question how other word length determining factors might possibly be integrated into the above quantitative analysis.

Finally another probability distribution

$$\alpha(m) = \frac{T}{m^{\nu}}\left(\frac{\lambda}{\mu}\right)^{m-1},$$

which is derived from a birth and death process as a stationary distribution, is fitted to the data. The denpendency of parameter values on word lenght is examined. Again we obtain different patterns for Finnish. We will discuss different attempts of explanation.

References:

CHITASHVILI, R. J. & BAAYEN, R. H.(1993): Word frequency distributions of texts and corpora as lage number of rare event distributions; in: G. Altmann & L. Hřebíček (ed.): Quantitative Text Analysis; (QL52); wvt: Trier 1993.

RAPOPORT, A.: Zipf's Law Re-visited; in: H. Guiter & H. Arapov (ed.): Studies on Zipf's Law; (QL 16), Brockmeyer: Bochum 1982.

Frequency Spectra within Word Length Classes

Edda Leopold

FB II; Linguistische Datenverarbeitung
Universität Trier
54286 Trier
Germany
Tel.: 0651/24625
e-mail: leopold@ldv35.Uni-Trier.de

Zipf-Mandelbrot law is applied to classes of words of equal length. The frequency spectrum of Zipf-Mandelbrot law is given by

$$\alpha^*(m) = \frac{1}{m^\gamma} - \frac{1}{(m+1)^\gamma} , \qquad (1)$$

where m denotes word frequency and γ is a positive parameter (see e.g. Chitashvili & Baayen 1993). According to the mean value theorem of differential analysis for each m>0 there is a $\xi_m \in [m;m+1]$ which satisfies

$$\frac{1}{m^\gamma} - \frac{1}{(m+1)^\gamma} = \frac{1}{\xi_m^{\gamma+1}} . \qquad (2)$$

The distribution

$$\alpha(m) = \frac{T}{m^{\gamma+1}} \; ; \; \gamma > -1, \; T^{-1} := \sum_{m=0}^{\infty} \frac{1}{m^{\gamma+1}} \qquad (3)$$

is therefore applied to the frequencies of words with equal length. Using data from the Celex database for English, Dutch, and German, we obtained a satisfactory fit for all classes of length. Parameter γ, however, is variing with word length.

English, Dutch, and German show a similiar pattern. Parameter γ takes on negative values for short words (3 or 4 letters). Its value increases monotonously as the word length is increased, crosses the zeroe line and finally reaches a constant level ($0<\gamma<1$).

Negative values of γ are in contradiction to the Mandelbrot's derivation of the Zipf-Mandelbrot law (see Rapoport 1982: 9f). Correspondingly equation (1) does not yield a proper probability distribution for $\gamma < 0$, in contrast to equation (3). The negative values of γ can be explained by the fact that words of equal word length do not differ much in effort of articulation. However this cannot explain the Finnish Data from the Oulu-corpus, which reveals a completely different behavior of the parameter γ.

PLENUM III

The Czech Language Institute
Academy of Sciences of the Czech Republic
118 51 Letenská 4
Prague, Czech Republic

e-mail: uhlir@feld.cvut.cz

Linguists vs the public: An electronic database of letters to the Language Service as a source of sociolinguistic information

Ludmila Uhlířová

The paper deals with language behaviour of language users in a concrete sociological role - in the role of questioners, or correspondents of the Language Service, The Czech Language Institute, Prague. Some features of their behaviour are modelled and motives for their behaviour analyzed.

project note

sociolinguistics, language culture

1. Language Service as a dialogue

The topic of this paper belongs to the sociology of language. It deals with language behaviour of language users in a concrete sociological role - in the role of questioners, or correspondents of the Language Service, The Czech Language Institute, Prague. I shall attempt to model some features of their behaviour, using methods of mathematical statistics and probability theory, and to find motives and reasons for their behaviour.

Language Service is an institutionally organized form of language treatment. The main task of Language Service, as it has been performed in many European countries already for decades, is to answer questions about language asked by the speakers of a language. Anybody who has a language problem and wishes to get some information about his/her mother tongue, may call, write a letter or make a personal visit to the Language Service of a

particular country. He/she may ask, e.g., how to spell or pronounce a word, what is the meaning of a word, and many other questions. The public in many countries are used to their Language Service and appreciate it as a kind of a cultural service.

I consider the Language Service to be a dialogue between linguists and the public. The public ask, and the linguists answer. But, at the same time, linguists also receive information. My point is that enquiries from the public represent a valuable source of sociolinguistic information about actual problems of language use, as well as about users themselves and about their attitudes to language. The importance of this information lies in the fact that it has sometimes immediate, sometimes long-term feedback effects on decisions of linguists as language managers.

2. The database

The data are taken from an electronic database of questions of the public and answers to them. The database was started in the Language Service, The Czech Language Institute, Prague, in 1992 and its size has been permanently growing. Each database record contains information of several types.

3. The sample data

From the package of information in the database I have chosen two items which I am going to discuss in more detail:

(a) social/professional groups of enquirers

(b) topics of queries.

The question to be answered sounds: What can be said about the mutual relationship between a and b?

4. Statistical processing

Statistical processing of the data offers preliminary

answers to several issues:

4.1. If a letter arrives (or a phone call is made) in the Language Service, then what is the probability that its sender or questioner is a person/institution from a particular social/professional group? (The task is to find a probability distribution which will properly model the observed distribution of values = number of queries from various social or professional groups).

4.2. Is there an association between the appurtenance of questioners to one of the ten social/ professisonal groups of questioners and a typical or less typical topic of their queries? The association/dissociation is tested. A contingency table is created for this purpose.

4.3. To what extent are the linguistic interests of each of the questioners' group concentrated on one or two topics, and to what extent are they dispersed? As a measure of the concentration of questioners' interests Herdan's repeat rate has been used.

4.4. What are the language interests, or problems of an average questioner? (What is the linguistic justification of the notion of "average questioner"?)

The statistical analysis of our data is still tentative and much more work is needed to make it more reliable. Still I hope that the results achived so far have shown that, in principle, an electronic database of the Language Service documents, is an effective tool not only in consulting practice, but also in sociolinguistic analysis. It is relevant for linguists to know the priorities of speakers' interests and problems, as manifested in their questions to the Language Service.

Multidimensional Scaling Methods Applied to a Computerized Dialect Atlas of Finnish

Sheila Embleton and Eric S. Wheeler
York University and Wheeler and Young Inc.

Address for correspondence:
Professor Sheila Embleton
Associate Dean, Arts
York University
4700 Keele Street
North York, Ontario
CANADA M3J 1P3
embleton@yorku.ca

Submission for "short paper" (15 minutes + 5 for questions)

Project note
Application of method (multidimensional scaling) to quantitative study of dialects (dialectometry) with special reference to Finnish

SUMMARY

Earlier research into the use of multidimensional scaling for presenting — and revealing the patterning in — complex or extensive dialect data has proven promising. Such studies require a machine-readable data source. For Finnish, the principal source is an out-of-print dialect atlas (Kettunen, 1940), which we are now putting into machine-readable form. Issues we consider include: data entry, error estimate, translation of typography, generality of data formats, intellectual property rights, and availability of original data sources.

1. QUANTITATIVE STUDIES OF DIALECT MATERIAL

We have been exploring the use of a computer-implemented statistical technique, called multidimensional scaling (MDS), for transforming detailed dialect information into more comprehensible pictures that resemble geographic maps (Embleton, 1993; Embleton & Wheeler, 1994, 1996). MDS is a useful (and increasingly used) technique in linguistic application; see for example Jassem & Lobacz (1995).

Our work until now has been on English dialects, employing an existing (although very recent) computerized version of the standard work on English dialects, the *Survey of English Dialects*. This computerized database was graciously provided by Prof. Dr. Wolfgang Viereck, of the University of Bamberg in Germany. The data provided, and we used, the linguistic choices on a list of 169 phonological, morphological, syntactic and lexical items, of speakers at 313 sites in England, to generate more than 50,000 linguistic "facts". These facts were transformed by MDS into a two-dimensional map in which sites with similar linguistic patterns are near one another.

No geographic information (such as map coordinates or latitude and longitude) is involved. Still, the resulting maps strongly reflect the actual geographic location of the sites, which is what we would expect from our understanding of the correlation between geography and dialects. The surprise is not from the correlation, but from the fact that we can see the correlation on one sheet of paper.

There are also relationships that contradict geography, thus highlighting where factors other than geographical distance must be at work. For example, on our maps of England, the south-west is closer to the north-east than it geographically is. But these are also the interesting observations for dialectologists to explain, for example by social, political, or historical factors.

Some of the relationships are artifacts of the process. The Isle of Man, remote from everywhere as a dialect, still ends up too near one place or another however we draw the map. We are looking at three-dimensional maps as a way of spotting these "outliers".

Finally, we note that MDS alone did not generate very readable maps. We had to use techniques to label and colour the maps so that the relationships in the map became more obvious.

2. THE CHALLENGES

Our studies came after some earlier work (e.g., Dobson & Black, 1979; Embleton, 1987a, 1987b; Shalit, 1984) showed that MDS worked on small amounts of data. Dobson & Black dealt with several Australian languages, and used their linguistic similarities to build a map roughly reflecting their geographic distribution, with any deviations explainable by borrowing patterns between languages. Shalit (1984) and Embleton (1987a, 1987b) all involved small pilot-studies using English dialect data.

But models that work on simple cases can still fail on cases with larger volumes of data. Our work with English dialects was partly to determine how well we could make the MDS technique apply to the volume of data — tens of thousands of data points — that normally come with a national dialect atlas. With suitable adaptations and extensions, we were able to get satisfactory results.

English dialects, however, are a well known situation. The macro-level dialectology has been studied extensively, the English data was already computerized, and the form of the English data was familiar to us as English-speakers. Would the technique extend to other languages, where perhaps that data is not computerized, the macro-level dialectology is less well-studied, the linguistic history is quite different, and the linguistic data is less familiar to us?

Finnish offers such a test case. The data has been extensively and reliably collected as well as published in atlas form. This standard dialect atlas (Kettunen, 1940) is widely referenced, but exists in its complete form only in out-of-print paper copies (which are much valued by their owners), although a subset of the data (exhibiting some of the most important dialect features) has been reprinted in a much smaller book; neither the full nor the "reduced" version of the dialect atlas is machine-readable.

The Finnish dialect situation is perhaps better understood at a micro-level than as a whole. In any case, this non-Indo-European (Uralic) language is superficially quite different than English. In addition, the volume of data was several times larger than the English data set we used. If there are methodological problems to encounter, we should find some of them here.

3. THE COMPUTERIZED DIALECT ATLAS OF FINNISH

To facilitate the scholarly study of Finnish dialects (by us and others generally) we are creating a computer-readable data set for the existing *Dialect Atlas of Finland* which we call the *Computerized Dialect Atlas of Finnish* (*CDAF*).

There are 213 maps and 530 sites at which data has been gathered. Each map has up to 16 (typically 4-8) features, giving us up to 1.8 million dialect "facts" for the Finnish data, or up to 36 times more data than the English data set (which had roughly 53,000 "facts").

Work Plan

Our plan of work is as follows:

Data acquisition

We have made photocopies of the out-of-print Kettunen atlas, from sources in Finland. These photocopies have been made in a considerably enlarged version, to make them easier to read, and hence the data entry less tiring and more accurate.

We have the appropriate permission from the copyright holder, Suomalaisen Kirjallisuuden Seura (SKS [Society for Finnish Literature]), not only to use and photocopy the materials, but also to publish any materials resulting from this research, including the distribution of the resulting computerized database, either on our own or in conjunction with SKS.

Data set design

The data set format and layout have been selected to be useful beyond our own application. It is largely based on what Viereck used for the computerized *SED* data, with some modifications and adaptations based on our experience with his format in our previous project.

Utility programs to facilitate fast and accurate data entry have already been written and tested.

Data transfer

A trained staff of student assistants will read the hard copy of the atlas, and enter data into a data collection program. Work methods will be rigorously defined, as consistency of data entry is paramount. Work will be spaced out over several months to minimize inconsistencies from fatigue and loss of focus.

Verification

Data verification steps need to be a part of the work assignments. Not only do we want the work to be very accurate because we expect this will become a standard reference for others to use, but we also want to express a numerical measure of what we believe the upper bound on residual error is. This is a non-trivial problem. Our current plan is that we will test random samples of the data, and compare for discrepancies, rather than doing double-data entry. Better would be double-data entry, because this provides a check of every entry. However, double-data entry is likely to be prohibitively costly in person-hours.

There will inevitably be rework resulting from the data verification step.

After completion, the data will be used in our ongoing research, particularly (but not limited to) MDS-related research on Finnish dialects.

Presentation for Research Use

We expect the data to fit on a small number of diskettes. It will be packaged with a description of how to use it, and how it was prepared.

The diskettes will be available for distribution at cost, from us directly as well as from SKS in Helsinki. We will only ask that proper acknowledgment be given, and that users report any errort to us, for correction.

Although the project is not yet completely funded, we have been able to acquire the base data, design the data sets and confirm the feasibility of our approach by doing some data entry.

4. ISSUES AND CONCERNS

Although computerizing a published data set is intellectually straightforward, there are a number of practical issues to face:

Software

The use of custom software applications for data entry adds to the project the complexity of ensuring the software works correctly, and functions in the various computer environments we have. The alternatives would be to use:

A standard database package. The database still needs to be designed and implemented, but the behind-the-scenes programming is done for one. However, one is then restricted to the capabilities and data storage formats of that package.

A simple data entry scheme, such as using a word processor to create the data files. Here, the risk of making errors in formatting is much greater than in a customized system that only permits acceptable entries.

For us, the custom application was the appropriate choice, because we had the ability to create it, and it gave us flexibility in shaping the output, as well as control over the

performance of those doing the input. In other circumstances, though, other choices would have been acceptable.
The accuracy of the data entry is a concern when the body of data is as large as this, and the odds are that even the most careful of data entry people will make some errors. We address this situation through a combination of "process control" and "statistically-based testing".

Process Control
For each printed map, the data entry application presents the data entry person with each site name in sequence and a list of only the valid dialect features for the current map. The person selects (with a mouse click) the right choice of feature for that location, and is moved to the next site name. Although it is easy to move back and forth in the data to review and revise, normally the data entry person has only to concentrate on the printed map.

Statistically-Based Testing
As part of reviewing and revising the data set, we intend to count the number of observed errors. Through standard statistical techniques, we can predict how many residual errors are in the data set, and adjust our search accordingly. As a bonus, we will be able to publish an upper bound on errors in the data set, so that other researchers know what they are using. As mentioned above, we will also expect all users to report any errors that they discover to us.

Data Format
The format of the computerized data needs to be such that a variety of researchers will be able to use it. Of concern to us is:
Having the data usable over a long period of time (decades or longer, as this is a standard reference work), and therefore probably outliving any current software package or database.
Having the data accessible to researchers around the world, who can be expected to use a wide range of computing technology.
Our strategy is to keep the data layout simple, avoid proprietary or highly specialized formats, and be prepared to reformat data if circumstances demand.

Special Characters
Finnish uses a slightly different orthography from English, and in addition the Kettunen Dialect Atlas has a variety of typographic characters and even some hand-drawn ones to represent particular dialect features. We have made some substitutions in representing these marks with the character sets available to us online, but in general we have been able to deal with the situation in ways that will be transparent to other users and without ad hoc solutions.

5. OUTLOOK

We expect that the creation of the Computerized Dialect Atlas of Finnish will take at least until 1998 to complete. At that time, we hope to have a tool for use not only in our own studies of quantitative methods for dialectology but for a wide range of scholarly studies in dialectology and in Finnish.

We will be pursuing our investigations into multidimensional scaling and other related work on making large data sets comprehensible and manageable. It is our hope that others will also be able to use the growing power of information technology and this computer-readable data to further their investigations of language, dialect, and Finnish.

REFERENCES

Dobson, Annette J., & Black, Paul. (1979). Multidimensional scaling of some lexicostatistical data. *Mathematical Scientist* 4, 55-61.

Embleton, Sheila. (1987a). A new technique for dialectometry. *Twelfth LACUS Forum*. Lake Bluff, Illinois: LACUS. Pages 91-98.

Embleton, Sheila. (1987b). Multidimensional scaling as a dialectometrical technique. In: Babitch, Rose Mary (ed.) *Papers from the eleventh annual meeting of the Atlantic Provinces Linguistic Association*. Pages 33-49.

Embleton, Sheila. (1993). Multidimensional scaling as a dialectometrical technique: Outline of a research project. In: Köhler, Reinhard & Rieger, Burghard (eds.) *Contributions to quantitative linguistics*. Dordrecht: Kluwer. Pages 267-276.

Embleton, Sheila, & Wheeler, Eric S. (1994). *Dialect project: Technical report.* York University, Toronto, Department of Languages, Literatures & Linguistics.

Embleton, Sheila, & Wheeler, Eric S. (1996). Multidimensional scaling and the SED data. In: Viereck, Wolfgang (ed.) *The Computer Developed Linguistic Atlas of England 2*. Tübingen: Max Niemeyer. (to appear)

Jassem, Wiktor & Lobacz, Piotra. (1995). Multidimensional scaling and its applications in a perceptual analysis of Polish consonants. *Journal of Quantitative Linguistics* 2, 105-124.

Kettunen, Lauri. (1940). *Suomen murrekartasto* [*The Dialect Atlas of Finland*]. Helsinki: Suomalaisen kirjallisuuden seura.

Shalit, Ami. (1984). Multidimensional scaling of linguistic distance between dialects (A computerized data analysis of dialect boundaries and transition areas). MA major research paper, University of Toronto, Department of Linguistics.

An explorative method concerning word classes

Gejza Wimmer, Gabriel Altmann

Abstract

For any classification of linguistic entities two very general hypotheses can be set up:

(1) If the classes of the taxonomy are ordered according to decreasing frequency then we obtain a regular rank-frequency distribution. There are many corroborations of this hypothesis (Zipf's law, Zipf-Mandelbrot's law, diversification laws, etc.) having many different forms.

(2) The classes of the taxonomy are in interaction, i.e. they compete or cooperate with oneanother. This hypothesis is corroborated by the whole qualitative linguistics whose main problem is to show the interaction of some classes (cf different kinds of grammar). However, in qualitative linguistics they have a rather deterministic form. From the quantitative point of view it is possible to develop methods enabling us

(i) to show *statistically significant* interrelations of classes and to display them graphically,

(ii) to characterize languages, genres, styles, texts quantitatively and/or graphically and

(iii) to compare these entities statistically by means of confidence intervalls or tests.

In the present paper we want to show a simple method using the linear model for solving problem (2i) as applied to classiccal word classes. The method, if applied to different languages or texts, can show whether the classification is stable or simply ad-hoc, i.e. whether it is adequate for specific purposes (general linguistic, grammatical or stylistic).

The method is exemplified on word classes of German and will be further developed.

STRUCTURE OF LANGUAGE - A QUANTITATIVE APPROACH

SIBASIS MUKHERJEE

LANGUAGE DIVISION, CALCUTTA. 234/4 A.J.C.Bose Rd.

17th.Floor, Nizam Palace, Language Div., Calcutta-20.700 020.

India. Tel. 91-33-240 0906. Fax.033-91-247 9926.

Summary : The relationship between the grammar and the lexicon on the level of vocabulary statistics is an interesting phenomena and also the field is challenged by many of the scholars. The present paper shows how structuralism can help to view the matter through quantitative linguistics.

The significance of the quantification concerning the linguistic structure, lies on the degree of the independence between the linguistic symbols and the semantic content of it. This independence was observed by de Saussure in earlier fifties which was later on developed by Trubetzkoy and his followers into the system of the phonological oppositions (privitive, gradual and equipollent) in Prague school. This was the structuralism in phonetic level. Now, going back to de Saussure's conception of language, the signifier and the signifie' are inevitably required for the formal analysis of language.Based on his idea, the total conception of structuralism has arisen which still requires further modern linguistic thoughts.

Coming back to Trubetzkoy's principle of phonology and the independence of linguistic form, the later is actually the case between the grammatical form and the particular lexical item. The grammar deals with the general facts of the language where as lexicology with the particular. As per Jesperson, `rat' denotes that particular animal is a special fact which concerns that word alone, but the formation of the plural by adding the sound -s is a general fact because it concerns a great many other words as well : plays, bats, books etc. But

this is also not totally correct since any lexical item cannot be used for every grammatical form.

Now, if we take quantitative linguistics as extension of structuralism to the vocabulary and syntax, for a statistical parameter there must exist a certain degree of independence between the linguistic symbols as stated earlier. On the phonemic level, structuralism is a matter of independence between the sound and the meaning of individual word and on the vocabulary level, it is the matter of independence between the frequency distribution of the linguistic forms and the content of the literary text.

Now, taking any literary text and arranging the vocabulary used once, twice ..... n times, we can come to know the distributional pattern of the classes in which the words used once (the largest class), twice (the second largest class), thrice etc. and so on. In this way we can reach the vocabulary items of the less general and less frequent class. So, we can find the items more frequent i.e. more general in use and also the items less frequent and less general use. It may vary from text to text. This is the general pattern, no matter what is the content of the text. Since continuous texts differ in the grammatical arrangement of the vocabulary items as well, the statistical concept of the use of the words can be regarded as the higher degree of structuralism. This is the most interesting fact that how Trubetzkoy extended the fundamental principles of Phonology. He has seen it as the device of the distinctive function of the phonemes in terms of the lexical items used for the purpose of grammar. Now, coming to the frequency of occurrences of the lexical items, the advent of Stylo-Statistics is remarkable. The formulation of the basic relationship between the vocabulary frequency and the style of a certain language

remains same if we exchange the `style' for `language' and vice versa. Therefore, on the phonemic level the elements having the distinctive functions or on the lexical level the elements of frequency help us to consider the style and the language as one.

The important fact here is to be mentioned is that, Yule (1944) had first attempted systematically to analyse the stylistic role of frequency where the frequency of use was claimed to be a style of a literary creation. He has a characteristic denoted as `K' denoting language as well as stylo-statistic parameter. But Ellegard's distinctiveness ratio characteristics is somewhat different. If we consider W stands for style , , we can get the following equation -

$$f_{ij} = \text{style } (j) / \text{language } (i)$$

$$f_{i1} = \text{style } (1) / \text{language } (i)$$

And from the above equation we get,

$$W_j = f_{ij} / f_{i1} \quad W_1$$

Analysing this, we get, the difference in style, because the word frequency between the sample and that of a particular writer is the frequency for a word in the vocabulary independent of the particular vocabulary item.

Therefore the distinction between the grammar and the lexicon on the level extended upto the vocabulary statistics is essentially one between the general and the special facts of the language. And it is the vocabulary statistics which creates the differences between the language and the style (in terms of the frequency elements) and thus the present paper envisages to highlight how struturalism can analyse the fact through quantitative view point.

FORMS

Dr. Karl-Heinz Best, Georg-August-Universität Göttingen, Seminar für deutsche Philologie, Humboldtallee 13, D-37073 Göttingen. E-mail: kbest@gwdg.de.

Topic: Theoretical linguistics
Project note

Zum Stand der Untersuchungen zu Wort- und Satzlängen

0. Summary

This paper presents the project on the distributions of word length and sentence length carried out mainly at the universities of Bochum and Göttingen. The project aims at discovering the laws controlling the frequency distributions of these and, perhaps, further entities. Up to now, more than 30 languages have been investigated with promissing results.

1. Vorbemerkung

Seit 1993 werden in einem Projekt zur quantitativen Linguistik, das in Kooperation mit G. Altmann (Universität Bochum) an der Universität Göttingen koordiniert wird, Daten zur Verteilung von Wortlängen in möglichst vielen Sprachen (bisher über 30 Sprachen) erhoben. In Ergänzung dazu wurden inzwischen auch Satzlängenverteilungen in ca. 250 deutschen Texten durchgeführt. Dieses Papier soll einen Überblick über den Stand der Arbeiten vermitteln.

2. Theoretische Grundlagen

Die theoretischen Grundlagen der Untersuchungen zur Wortlängenverteilung wurden in Fucks (1955/ 1955a), Grotjahn (1982), Wimmer/ Altmann (1996) sowie Wimmer u.a. (1994) gelegt; in diesen Arbeiten wurden folgende Vorschläge für die Verteilung von Wortlängen in Texten unterbreitet:
W. Fucks (1955/ 1955a):
⇒ verschobene Poisson-Verteilung (Annahme lt. Grotjahn [1982: 55]: "daß die einzelnen Ereignisse voneinander unabhängig sind und mit einer konstanten Wahrscheinlichkeit auftreten.");
R. Grotjahn (1982):
⇒ verschobene zusammengesetzte Poissonverteilung (= verschobene "negative Binomialverteilung" [Grotjahn 1982, 57f.] Annahme: "Es dürfte jedoch weit eher der sprachlichen Realität entsprechen, wenn man annimmt, daß zwar jedes einzelne Wort einer verschobenen Poisson-Verteilung folgt..., daß jedoch die Wahrscheinlichkeit nicht für jedes Wort gleich ist, sondern in Anhängigkeit von Faktoren wie (sprachlicher) Kontext, Themawechsel etc. variiert. Dies bedeutet,

daß der Parameter Θ der verschobenen Poisson-Verteilung selbst wieder als Zufallsvariable anzusehen ist" [Grotjahn 1982: 55]. Durch Einsetzen einer Gammaverteilung für Θ kommt Grotjahn zur zusammengesetzten Poisson-Verteilung als Modell der Wortlänge.).
G. Wimmer u.a. (1994: 101) gehen von der Annahme aus, daß die Wortlängenklasse P_2 in Texten proportional zur Wortlängenklasse P_1 erscheint:
$P_2 \sim P_1$.
Nimmt man an, daß zwischen den Wortlängenklassen kein konstantes Verhältnis herrscht, kann man entsprechend ansetzen:
$P_x = g(x)\, P_{x-1}$.
g(x) kann verschiedene Formen annehmen; je nachdem, welche spezielle Form die Funktion annimmt, kommt es zu
⇒ einer ganzen Gruppe von Funktionen, darunter die Hyperpoisson-V., die Hyperpascal-V., die negative Binomialverteilung oder die Palm-Poisson-Verteilung.
G. Wimmer/ G. Altmann (1996) erweitern dieses Konzept zu einem System unter Einbeziehung weiterer Funktionen.
Die beiden zuletzt genannten Arbeiten bilden den theoretischen Hintergrund der bisher durchgeführten Untersuchungen.

3. Zum Stand der Untersuchungen

Es werden möglichst viele Texte aus möglichst vielen Sprachen bearbeitet. Eine Streuung nach Zeit, Textsorte und Autor wird angestrebt.
Bearbeitungsprinzipien: vgl. Altmann/ Best (1996).

3a. Deutsch

Gegenstände:

Ahd.: überwiegend literarische Texte;
Mhd.: literarische Texte und Rechtstexte;
Fnhd.: Briefe, Tischreden;
Nhd.: Briefe, viele Pressetexte der neuesten Zeit; Kurzprosa. Andere Textsorten (z.B. Gedichte) nur sporadisch.

Ergebnisse:

a) ahd., mhd.: Poisson-Verteilung; auch Hyperpoisson-V., bei der die Ergebnisse insgesamt etwas schlechter ausfallen. Problem: Die Hyperpoisson-V. hat einen Parameter mehr als die Poisson-V. und ist deshalb etwas weniger flexibel bei der Anpassung an Texte mit sehr wenigen Wortlängenklassen.

b) Briefe (fnhd. bis heute): Hyperpoisson (nur 1 Brief nicht!)
c) andere Textsorten: Barockgedichte: alle Hyperpoisson
Pressetexte: pos. negat. Binomialverteilung (z.T. auch Hyperpoisson)
Pressetexte mit stark fachsprachlichem Einschlag: gemischte Poisson-V., positive Singh-Poisson-V.
Kurzprosa: Hyperpoisson-V., gem. Poisson-V., pos. neg. Bin.
Problem: Es sind noch nicht bei allen Textsorten alle Verteilungen geprüft.
Briefe: ideale Textsorte, da spontan verfaßt, kaum überarbeitet, nicht zu lang,...d.h., relativ homogene Textsorte.

3b. Niederdeutsch

Fast an alle 110 Texte (49 Alltagssprache, 31 Kurzgeschichten, 30 Gedichte) läßt sich die positive negative Binomialverteilung anpassen. (Alltagsverkehr: 2 nicht, dafür andere Modelle; Kurzprosa: 2 mal schwach; Gedichte: 2 mal nicht).

3c. Fremdsprachen

Zum Datenbestand gehören z.Zt. über 30 Fremdsprachen. Zu einer davon, Irisch, wurden noch keine Berechnungen durchgeführt, da nach Auskunft des Bearbeiters erhebliche Entscheidungsprobleme bestanden und mit nur 8 Texten eine unzureichende Datengrundlage vorliegt.
Relativ gut belegt: Nordgermanisch, Englisch
Griechisch (alt und neu)
Latein und romanische Sprachen
Slawische Sprachen
Finnisch-ugrische Sprachen
Türkisch
Chinesisch
Koreanisch
Weniger gut (immer nur 1 Datensatz):
Althebräisch
Eskimo
Japanisch
Ketschua
Maori

Ergebnis: Bei alten Sprachen und bei vielen gegenwärtigen Sprachen bewährt sich immer wieder die Hyperpoisson-Verteilung (Latein, Althebräisch, Altgriechisch, Altisländisch, begrenzt Althochdeutsch; aber auch bei so divergenten Sprachen wie Neuhochdeutsch, Ketschua, etlichen finn.-ugr. Sprachen, Eskimo).
In andern Fällen scheint sie nicht infrage zu kommen, z.B. Chinesisch, Finnisch...
Tendenz: In vielen Fällen zeigt sich, daß ein Modell, das sich an eine Textsorte

anpassen ließ, auch ein gutes Modell für andere Textsorten und Zeitabschnitte der gleichen Sprache darstellt.
Bisweilen zeigt sich aber auch ein Einfluß von Textsorten, so z.B. bei Pressetexten im Deutschen mit vs. ohne starken fachsprachlichen Einschlag.
Probleme bei der Anpassung von Modellen gab es bei lappischen und chinesischen Pressetexten, letztere mit stark fachsprachlichem Einschlag; chinesische Texte mit stark fachsprachlichen Zügen zeigen einen deutlich oszillierenden Datenverlauf.

4. Perspektiven und offene Fragen

Nachdem zu allen schriftlich überlieferten Entwicklungsphasen des Deutschen inzwischen wenigstens eine Textgruppe untersucht ist, wird als Nahziel angestrebt, für einige davon eine breitere Streuung der Wortlängenverteilungen nach Textsorten zu erreichen. Außerdem sollen möglichst noch weitere Fremdsprachen zusätzlich erarbeitet werden, besonders solche, die zu typologisch bisher nicht berücksichtigten Gruppen gehören. Für die besser bearbeiteten Sprachfamilien ist an Untersuchungen zu historischen und strukturellen Zusammenhängen zu denken.

Als offene Fragen sind derzeit zu nennen:
a. Welche Rolle spielen die Definitionen der für die Untersuchung relevanten Einheiten? Bei englischen Texten wurde z.B. mal mit, mal ohne Triphthonge gearbeitet, ohne daß dies zu erkennbaren Divergenzen geführt hätte. (z.B. "fire": 1- oder 2-silbig?)
b. Wie behandelt man nullsilbige Wörter am besten? Sollte man sie als selbständige Wörter behandeln oder wie Enklitika als Bestandteile ihrer Nachbarwörter? Bei der Anpassung der erweiterten positiven Binomialverteilung an tschechische Texte konnte beobachtet werden, daß bei einer Gruppe von Briefen die Anpassungen an Dateien ohne nullsilbige Wörter wesentlich besser waren als die an Dateien mit nullsilbigen Wörtern.
c. Sind die gefundenen Modelle immer die theoretisch und empirisch besten?
d. Lassen sich Fortschritte bei der Interpretation oder gar Prädiktion der Verteilungsparameter erzielen?

5. Weiterungen

a. Ausgehend von der Annahme, daß die Längen anderer sprachlicher Größen sich möglicherweise ähnlich verhalten könnten wie die Wortlänge, wurden bisher auch 3 umfangreiche Untersuchungen zur Satzlängenverteilung in deutschen Texten durchgeführt: An fast alle Texte läßt sich die Hyperpoisson-Verteilung anpassen.
Sehr naheliegend ist hier die Frage nach der Längenverteilung anderer Größen wie Silben, Morphe, etc.

b. Eine weitere Frage im Zusammenhang mit den bereits erhobenen Daten besteht darin, ob man diese nicht zu anderen, bisher nicht berücksichtigten Zwecken nutzen kann. Es bietet sich an, sie z.B. für die Entwicklung eines Syntheseindex in der Sprachtypologie einzusetzen.

Literatur

Altmann, G./ Best, K.-H. (1996). *Principles for Word- Length Count.* Unpublished Paper

W. Fucks. (1955). *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen.* Köln/ Opladen

ders. (1955a). Theorie der Wortbildung. *Mathemat.-physikal. Semesterberichte 4.* 195-212

R. Grotjahn.(1982). Ein statistisches Modell zur Verteilung der Wortlänge. *Zeitschrift für Sprachwissenschaft 1*: 44-75.

G. Wimmer/ G. Altmann. (1996). The Theory of Word Length Distribution: Some Results and Generalizations. *Glottometrika 15.* 112-133

G. Wimmer/ R. Köhler/ R. Grotjahn/ G. Altmann. (1994). Towards a Theory of Word Length Distribution." *Journal of Quantitative Linguistics* 1: 98-106

Bibliographie und weitere Informationen zum Projekt finden sich im Internet unter der Adresse: http://www.uni-goettingen.de/FB/Hist. Hier findet man das "Seminar für deutsche Philologie", direkt dabei unser "Projekt Quantitative Linguistik".

Adaptive Parameter Training in an Interpolated N-gram Language Model*

P. O'Boyle, J. Ming, M. Owens, and F. J. Smith

School of Electrical Engineering and Computer Science,
The Queen's University of Belfast
Belfast BT7 1NN
N. Ireland
UK

p.oboyle@qub.ac.uk

Summary

We show how an interpolated n-gram language model with adaptive parameters can be constructed. The model is tested using perplexity as a measure and compared with other similar models, which have fixed parameters; it is shown to perform as well as the best of these. The new model technique used to adapt the parameters is not computationally intensive and removes the need for a substantial amount of pre-training.

topical paper

topic area: statistical language modeling

Introduction

Interpolated n-gram models[1,2,3] are used in speech recognition to estimate the a priori probability of the various word sequences that may have been uttered in a particular instance. These language model probabilities are calculated as the product of a series of conditional probabilities for words following a given sequence of words. Ultimately these conditional probabilities are calculated from unigram, bigram, trigram, and higher order distributions derived from a large training corpus appropriate to the recognition task. The exact proportion of each distribution in the final mix is determined by parameters within the model; current models use a fixed set of parameters that are optimized during training and remain fixed during testing (or recognition). We introduce a new approach in which the parameters do not need to be pre-trained, but are continuously updated during testing of the model. This allows the model to adapt its parameter set to suitable values for each new test text.

* This research was supported by EPSRC grant GR/K82505

Our results show that the performance of the adaptive model is at least as good as that of the best existing interpolated model (with fixed parameters) but requires considerably less time to construct.

The n-gram model

Due to the sparse data problems inevitably encountered when processing natural language basic maximum likelihood estimates for conditional probabilities are unacceptable; words appear in new contexts even when the training data contains many millions of words. To overcome this problem interpolated models combine a number of maximum likelihood probability estimates to produce a single smoothed probability estimate. This can be achieved by first combining the two lowest order estimates and then incorporating other higher order estimates in turn until the final probability is obtained. In our experiments we have used this structure to combine five levels of probability estimates.

For convenience we use the following notation: w_i represents the *ith* word in a sequence of words and w_i^j represents the sub-sequence of words from the *ith* word to the *jth* word. Using this notation the interpolated model is defined by the following equations:

$$P(w_i|w_{i-1}) = \lambda_1 p_{ML}(w_i|w_{i-1}) + (1-\lambda_1) p_{ML}(w_i) \quad (1)$$

$$P(w_i|w_{i-j}^{i-1}) = \lambda_j p_{ML}(w_i|w_{i-j}^{i-1}) + (1-\lambda_j) P(w_i|w_{i-j+1}^{i-1}) \quad (2)$$

where λ_j are the parameters of the model, p_{ML} are maximum likelihood probabilities derived from a large training text, and P is the models probability estimate.

Adaptive Model

In our adaptive model no initial training is used to set the λ_j parameters (though suitable pre-training is possible). All parameters start with a default value and are updated following each use while processing the test text. This update is based on the iterative optimization used to train parameters in conventional fixed parameter models[4]. The value of the parameters are determined as a quotient of two values that are updated to modify the value of the parameter. Thus

$$\lambda_j = A_j / B_j \quad (3)$$

We now define the adaptive nature of the model by giving the formula used to update the A_j and B_j values.

$$A'_j = \delta A_j + \left(\prod_{k=j+1}^{n} (1-\lambda_k) \right) \lambda_j p_{ML}(w_i|w_{i-j}^{i-1}) / P(w_i|w_{i-n}^{i-1}) \quad (4)$$

$$B'_j = \delta B_j + \left(\prod_{k=j+1}^{n} (1-\lambda_k) \right) P(w_i|w_{i-j}^{i-1}) / P(w_i|w_{i-n}^{i-1}) \quad (5)$$

Equations (4) and (5) give the updates used in our adaptive n-gram model; a decay factor δ, which should be less than 1 is included in the model. Suitable initial values for A_j and B_j are selected to initialize the model. In these experiments the values 1 and 2 respectively have been used for these initial values to give λ_j an initial value of 0.5. It is possible to train these initial values; however, as we are more concerned with the ability of the model to adapt to the test text than with its ability to start from a best position, we use only the simple initialization described above.

We can extend the model described above to incorporate multiple λ_j values for each value of j. For example, when combining unigram and bigram estimates the choice of appropriate λ_1 value can depend on the frequency of the preceding word w_{i-1}. In this way bigram estimates based on a word with a high frequency can be given more significance than those based on a word with a lower frequency. We apply this principle to all our λ_j values by dividing the frequency range for j word phrase into a number of broad bands and maintaining a separate parameter for each such band.

Experimental details

The models have been tested using text taken from a large corpus of newspaper articles. Models trained using a deleted estimate and held-out training algorithms are also constructed for comparison with the new models using perplexity as a measure of performance. The test text has been selected as a number of large blocks distributed evenly throughout the text and was not used to train any of the models. The results therefore show the performance of the adaptive model where the performance of fixed parameter models are near to their best (i.e. when the test and training texts are very similar).

The parameters for the held-out model were trained using a second sample of test text from the corpus. This text was also used for some tests during the development of the adaptive model.

ML probability estimates are trained using 8,726,007 word tokens of the text with a vocabulary of 81,498 word tokens; words in the test text that do not appear in the training text have zero probability and are removed from the perplexity calculations (the test text contains 88,113 word tokens of which 701 need to be removed from perplexity calculations).

All parameters for the held-out and deleted estimate models are trained so that a minimum of 2,000 training examples are used to train each value. Based on this restriction the held out model has a total of 94 parameters and the deleted estimate model has a total of 4,770 parameters.

We have tested the adaptive model described above with a range of decay values and with different numbers of parameters. The number of parameters in the model is determined by setting an upper limit on the number of tokens that a single frequency band can contain.

The perplexity is inversely proportional to the (geometric) average word probability generated by the model. Thus a lower perplexity indicates a higher average probability and thus is generally thought to indicate a better performance.

Results

For comparison the held-out model produced a perplexity of 228.544 on the test set, and the DE model produced a perplexity of 227.050. Table 1 contains perplexities for the adaptive model with a range of sizes of parameter sets and decay values.

Table 1 Perplexities for an adaptive language model with a range of decay values and number of parameters.

No of Parameters	No Decay δ=1	δ=0.995	δ=0.99	δ=0.98	δ=0.96
65	227.248	227.021	227.038	227.270	228.206
34	227.241	226.942	226.871	227.018	227.888
17	227.658	227.054	226.874	226.949	227.729

The results in table 1 show that the dynamic model has produced a perplexity below that of the best fixed parameter model for a number of parameter set and decay value combinations.

Conclusions

We have shown that interpolated n-gram models can be produced without the need for pre-training of the interpolation parameters. The results in table 1 show that the performance of this model is as good as that of existing models with fixed pre-trained parameters. The best perplexity, 226.871, for the adaptive model is obtained with 34 parameters in a model and a decay factor δ=0.99. This is slightly better than the performance of the best fixed model, 227.050, which contains considerably more parameters.

The processing requirements of the models are also significant. The adaptive model can be used once the maximum likelihood probabilities from the training text are available, the held out model requires a small additional period for parameter training, and the deleted estimate requires a considerable additional period for training parameters. In our experiments the adaptive model and the held out model required a similar amount of training time while the training time for the deleted estimate model was an order of magnitude larger. The construction of the adaptive model is comparable to that of the weighted average model[1] but its performance is superior.

The perplexities in table 1 are all similar indicating that the performance of the adaptive model is relatively insensitive to the variations in the number of parameters and value of the decay factor. We expect a more significant improvement over fixed parameter models when the test text is less similar to the training text. We are also currently investigating models that dynamically change the size of their parameter set during testing which may lead to further improvements.

In these experiments we have used the adaptive parameter training technique to combine fixed maximum likelihood n-gram probability distributions; however, this technique can also be applied to combine other non static distributions, such as a cache based model. We believe this will be a particularly fruitful application for the dynamic adaptation of parameters, which can ensure that a near optimal mix of distributions is maintained.

References

[1] O'Boyle, P., Owens, M., Smith, F. J. "A weighted average ngram model of natural language", Computer Speech and Language, Vol. 8, 337-349, 1994.
[2] O'Boyle, P., Ming, J., McMahon, J., Smith, F. J. "Improving n-gram models by incorporating enhanced distributions", IEEE ICASSP'96, Vol. I, 168-171, Atlanta, 1995.
[3] Katz, S. M. "Estimation of probabilities from sparse data for the language model component of a speech recogniser", IEEE Transactions on Acoustics Speech and Signal Processing, ASSP-35, 400-401, 1987.
[4] Jelinek, F., Mercer, R. L. "Interpolated estimation of Markov source parameters from sparse data", In Pattern Recognition in Practice, editors Gelsema, E. S. & Kanal, L. N., 381-397, North Holland Publishing Company, Amsterdam, 1980.

LEXICON

LEXICON

Network analysis of vocabulary lists

Peter Forster

Heinrich-Pette-Institut für Experimentelle Virologie und Immunologie
an der Universität Hamburg, Martinistrasse 52, D-20251 Hamburg, Germany
Tel.: +49-40-48051-288 Fax: +49-40-48051-188
email: forster@hpi.uni-hamburg.de

Category: Preliminary version of a short paper (15 minutes + 5)

Topic: Multivariate combinatorial analysis in comparative linguistics

Summary
I unite the tree model and the wave model of language evolution into one geometric network model. This approach has previously been used for reconstructing DNA evolution, and I now apply it to vocabulary lists of closely related languages. The results mostly confirm traditional language relationships, but also offer interesting details.

Two well-known models to explain the mechanism of language evolution are the "tree model" and the "wave model" (reviewed by Goebl 1983). These are usually thought to be complementary rather than alternative mechanisms, and it would therefore be desirable to visualise both these aspects of language evolution in a single diagram.

This type of problem happens to be perfectly tailored to network methods that were originally developed for reconstructing phylogenetic relationships (which are not necessarily treelike) from DNA sequences (Bandelt 1994; Bandelt et al. 1995): During evolution, a given DNA sequence will acquire mutations at random positions, causing the progeny sequences to become more and more dissimilar from one another and from their ancestral sequence as time passes, yielding the "treelike" aspect of DNA evolution. Occasionally however, mutations at the identical position in two different DNA sequences will cause these two sequences to become more similar, resulting in convergence.

Extensive convergence is visualised as weblike reticulations in a phylogenetic network, and absence of convergence will shrink a reticulate network to a simple, unique shortest tree ("maximum parsimony" alias "Steiner" tree) . The phylogenetic network can thus represent both extremes of evolutionary outcomes, and has successfully been applied to questions on human evolution (e.g. Richards et al. 1996; Forster et al. 1996).

Which form of linguistic information is most amenable to phylogenetic analysis? I have chosen the 100 word list proposed by Swadesh (1955) to characterise the basic vocabulary of a given language. Word stem replacements as well as uptake of ancient idiosyncratic substrate contribute to divergence between originally related vocabularies (the "tree" aspect of language evolution), whereas convergence of vocabularies is caused by loan events of word stems (the "wave" aspect of language evolution). Therefore, although the mechanisms of DNA and vocabulary evolution are not analogous, the effects seem to be. Thus, the treelike as

well as the wavelike events of vocabulary evolution can be united and visualised in a phylogenetic network.

For the algorithm, it is furthermore important to note that no coalescence is necessarily assumed, as languages (in contrast to DNA) may have originated from independent sources. Incidentally, I do not employ the word lists for dating language splits, as Swadesh attempted. His approach appears to be flawed for several reasons, one of which is that he did not distinguish language substrate from superstrate.

I have collected the word lists from Indoeuropean languages based on personal interviews of at least two independent native speakers for every language in order to avoid the bias that is incurred when using dictionaries. In order to gain a representative sample of the variety within a given language family, I sampled all mutually unintelligible speech variants, which I defined as "languages" following the recommendation of Grimes (1988). Using this criterion, Bavarian German and standard High German for example, which are mutually unintelligible to unpractised native speakers, qualify as separate languages. Phylogenetic networks were then constructed from these word lists as described by Bandelt (1994). The networks yield an aesthetically appealing picture of language relationships in that the traditional language trees are mostly (but not always) confirmed, while the visualised loan word exchanges demonstrate a clear geographic pattern.

I then turn to the more interesting question of whether it is possible to discern from the network if (a) a given language family has evolved from an ancestral ursprache or (b) the relationships have evolved from exchange between originally distinct languages. I propose the following criteria:

(a) In the extreme case that all present similarity derives from a common ursprache, and all dissimilarity has evolved since, the phylogenetic relationships in a network should be dominated by the *migration routes* of the founding urvolk as well as by the (unknown) *time depths* of linguistic splits. This would yield a treelike network which may however contain a certain amount of reticulation due to loan exchange between neighbouring languages.

(b) If at the other extreme, present similarity within a given group of languages has arisen solely through extensive exchange of unrelated preexisting languages, the phylogenetic relationships should be dominated by *geographic proximity*, and ideally, the network should contain a *consensus language*.

The intermediate case between these extremes would be a blend of substrate, superstrate, subsequent loan exchange, and regional word stem replacement, and my analyses of Indoeuropean language families show characteristics of both outcomes, as may be expected from previous historical and linguistic research. However, a few languages in the network do not fall in line with traditional classifications.

Well aware that vocabulary only constitutes one aspect of language, I would encourage the development of similarly informative character lists for grammar and for phonetics to check if independent information can confirm these results.

References

Bandelt H-J (1994) Phylogenetic networks. Verh naturwiss Ver Hamburg NF 34:51-71

Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141: 743-753

Forster P, Harding R, Torroni A, Bandelt H-J (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59:935-945

Goebl H (1983) „Stammbaum" und „Welle". Zeitschrift für Sprachwissenschaft 2:3-44

Grimes BF (1988) Ethnologue. Languages of the world (11th edition). Summer Institute of Linguistics, Dallas, Texas

Richards M, Corte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, Hedges R, Bandelt H-J, Sykes B (1996) Palaeolithic and Neolithic Lineages in the European Mitochondrial Gene Pool. Am J Hum Genet 59:185-203

Swadesh M (1955) Towards greater accuracy in lexicostatistic dating. Internat J Am Linguistics 21:121-137

Rule-and network oriented approach to the semiotic model of the linguistic sign.

Authors: Valentin A. Chizhakowski, Anatol. N. Popescu

Addresses:
Valentin A. Chizhakowski: *str. Mircesti 8/1, ap.52; Chisinau 2049, Moldova.*
Anatol. N. Popescu: *str. M. Basarab 10/2, ap. 23, Chisinau 2045, Moldova. Email: anatol@accent.moldova.su*

Affiliation: Technical University of Moldova

Summary: There are presented the existing semiotic theories of the sign and our appreciation. We are proposing a new approach based on rule-and network oriented model of the sign. The latter are based on production rules and Petry networks theory and systems modeling.

The words: semiotic theory, linguistic sign, Petry networks, sign modeling, semiotic model

Specification of the logic are: Methodological problems of model construction.

1. Semiotic theories of the sign.

At present there exist 4 fundamental semiotic theories of the sign [10]: the logical one (G. Frege [8]), the linguistic one (F. de Saussure [7]), the pragmatic one (Ch.-S. Peirce, Ch. Morris, G. P. Melnicov [1,4]), the engineer-linguistic one (R. G. Piotrowski [5], E. A. Shingariova [10], V. A. Chizhakowski [9]).

We shall carry out a short analysis of the four fundamental semiotic theories and introduce the necessary definitions.

At G. Frege the sign is a material bearer of a notion (sense) referring to an object (denotatum), i.e. simply a material «label» (a notion mark) acting as an equal element of the logical triad «object-notion-sign».

According to F. de Saussure the sign is a double-sided psychological essence: a combination of the content level (notion, signified) and the expression one (signifying), i.e. the psychical unity of the signified and the signifying (the acoustic or graphic form of the signal).

According to the theory of Ch.-S. Peirce the sign is the result of the reflection, at the beginning of a dynamic object into a direct one (the ideal form of the object in the consciousness of the speaker) and after this direct object is confronted in the act of semiosis (sign formation) with the interpretanta (notion allowing different interpretations). In that way the sign is

defined both as the material substitution for an object of the real world in the limits of the logical triad «object-interpretanta-sign» and as the interpretanta itself.

The sign scheme proposed by R. G. Piotrowski will be taken by us as a base when elaborating our sign network model. On account of this we shall introduce some definitions for the component structure of the sign being used in the given theory [3].

Definition 1. Denotatum D_n represents an integral and not dismembered by our consciousness reflection (form) of an isolated referent (an object of the surrounding world) or a generalizes-typified form of an entire class of objects (generalized referent).

Definition 2. Designatum D_s represents a notion unit expressed through a concept (the main mark) or an intentional (a selective totality of marks), the essence of a certain referent (class of objects).

Definition 3. Connotatum C_n of a sign is a complex of traces of the second semantization generated as a result of metaphorical utilization of the sign, of their emotive or stylistic coloration connected with the preferential utilization of the sign in a definite variety of language (functional style, sublanguage, dialect or version of literary language).

By virtue of the introduced definitions denotatum reflects the extentional aspect (volume) of the notion correlating it to situations of the surrounding world [11]; designatum-the taxonomical form of the referent consisting of a collection of taxonomical (classificatory) marks of the notion indicating the place of the referent in those or others classes (taxons) of the classificatory network. Designatum reflects the intentional aspect of the notion content. Connotatum-the emotional form of the referent covering the emotive marks of the referent.

Definition 4. The referent *r* is an object of the real world signified in a sign.

Definition 5. The signal referent *Sr* is a chain of acoustical or graphical signals.

Definition 6. The noun *J* is the acoustical form of the material cover of the sign kept in the consciousness of a speaker.

We shall also offer you the semiotic model of the sign which is attributed to linguistic or communicative orientation going back to the theories of F. de Saussure and Ch.-S. Peirce and developed in the works of R. G. Piotrowski [2,3] and E. A. Shingariova [10].

Definition 7. The sign represents a double-sided psychical unity of the signified consisting of denotatum (D_n), designatum (D_s) and connotatum (C_n) and signifying (noun, J) correlated with the referent (*r*) and the signal referent (*Sr*).

The presented fundamental theories of the sign posses the following shortcomings [10].

1. The nature of the sign is described unilateral. So, for example, it is «simply a material label» (G. Frege), «a double-sided psychical essence» which is cut off from its material bearer (F. de Saussure), or a material substitutor of the object» and «interpretanta» (Ch. Peirce).

2. Each of the first three theories either describes one status of the sign Z(the semiotic, the linguistic or the communicative one) or doesn't draw boundary-lines between them. The semiotic status of the sign in introspection from the referential logic of the world is described by G. Frege; the same status but in introspection from the language is described by F. de Saussure. Ch. Peirce refers to the semiotic and communicative statuses without establishing boundary-lines between them.
3. All the four fundamental theories present the sign in the form of a tough static structure which doesn't allow to model the process of semiosis and construction of complex sign structures.
4. In all of the analyzed theories there isn't an accurate division between the sign models and sign formation structures used in the communication process for the primary and second semiosises.
5. In the engineer-linguistic theory of sign pragmatics which presents by itself «the combination of the given sign with a person» and is the valence of the sign [2] isn't carried explicitly through or is presented in the form of an external vector stretching from communicators to the sign and picking out in each component of the sign its informative kernel [10, p.15].

Now we shall demonstrate our new rule-and network oriented model of the sign with the help of which we intend to remove the above analyzed shortcomings.

2. Rule-and network oriented model of the sign.

Using positions and transitions [6] we shall construct a network which models the sign formation process. For this purpose we shall define the sign in its first approximation as a double-sided psicholinguistic essence consisting of a signified (meaning, psychological form of the referent) and of a signifying (noun, psychological form of the signal).

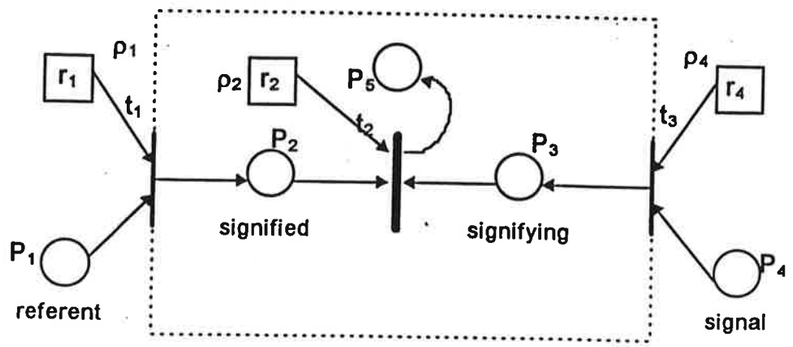


Fig. 1.

Definition 8. The static positional-transitional network (fig. 1.)used for modeling the sign formation may be presented by the four:

$$C_\varepsilon = (P, T, I, O),$$

where $P=P' \cup P''=\{P_1, P_2, P_3, P_4, \rho_1, \rho_2, \rho_3\}$ is a multitude of positions, $T=\{t_1, t_2, t_3\}$ is a multitude of transitions, $I:T\to P$ is the input function and $O:T\to P$ is the output function.

The position P_i contains referents, i.e. objects of the real world marked in the sign. We shall mark this multitude of referents $F=\{f_1, f_2, \ldots, f_i, \ldots, f_n, \ldots\}$. Each element (object) of the multitude F is being constructed from a multitude of subelements, the so-called attributes $\{f_{i1}, f_{i2}, \ldots, f_{in}\}$.

The signified which we shall present in the form of a reasoning space, i.e. of a field of interpretation

$M = \{\mu_1, \mu_2, \ldots, \mu_i, \ldots, \mu_n\}$ are transferring into P_2.

The same multitude but with the use of attributes is reduced to the form

$M = \{\mu_{ij}\}$, $i = \{1,n\}$; $j = \{1,n\}$.

The signal referents, objects in the form of acoustic chains or graphic signals, are entering from P_4 into the interpretator input. We shall mark them in the form of the multitude

$A=\{a_1, a_2, \ldots, a_i, \ldots, a_n, \ldots\}$

and the attributes which are to be included we shall present in the form $A = \{a_{ij}\}$, $i= 1,n$; $j= 1,n$.

The whole set which is used in a concrete natural language (NL) of signifyings will be considered as an alphabet

$$\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_i, \ldots, \alpha_n, \ldots\} \text{ or}$$

$$\alpha = \{\alpha_{ij}\}, \quad i=1,n; \quad j=1,n$$

and we shall allot for them the position P_3.

The psycholinguistic essence representing the signs will be presented in the form of the multitude:

$$\varepsilon = \{\varepsilon_i\} \text{ or } \varepsilon = \{\varepsilon_{ik}\}, \quad i=1,n; \quad k=1,k_i.$$

During the semiosis process they are transferring into the position P_5.

The transitions t_1 and t_3 which are realizing the reflection (interpretation) of the objects of the surrounding world in the consciousness of man will be called transitions of reflection and the transition t_2 - transition of semiosis.

The positions ρ_1, ρ_2 and ρ_3 are embodying the solvable functions r_1, r_2 and r_3 which are being used to realize the processes of reflection and sign formation.

For modeling the semiosis process we shall construct a dynamic network taking advantage of the following definition.

Definition 9. The dynamic network or the marked and painted C_ε-network which is modeling the semiosis process can be presented by the six $D_\varepsilon = (P, T, I, O, M, C)$, where P, T, I and O are correspondingly the multitude of positions and transitions, input and output functions;

M - is the function of marking $M:P\to N$, representing the vector $M=(M(P_1), M(P_2), M(P_3), M(P_4), M(P_5))=(m_1, \ldots, m_5)$ from the quantities of the markers of the material and

ideal objects (markers-out, counters, variables) correspondingly in the position P_1, P_2, P_3, P_4 and P_5; $C=\{C_1, C_2, C_3, C_4, C_5\}$ is the multitude of colours into which are painted the counters functioning in the network.

The sign formation process becomes possible only in the case if the dynamic network is included into the communication process, i.e. the given network is extended on the account of two pairs of the positions P_6 and P_7, ρ_6 and ρ_7 and of two transitions t_4 and t_5, related to the sender and the addressee. The dynamic network which is included into the communication act is presented in fig. 2.

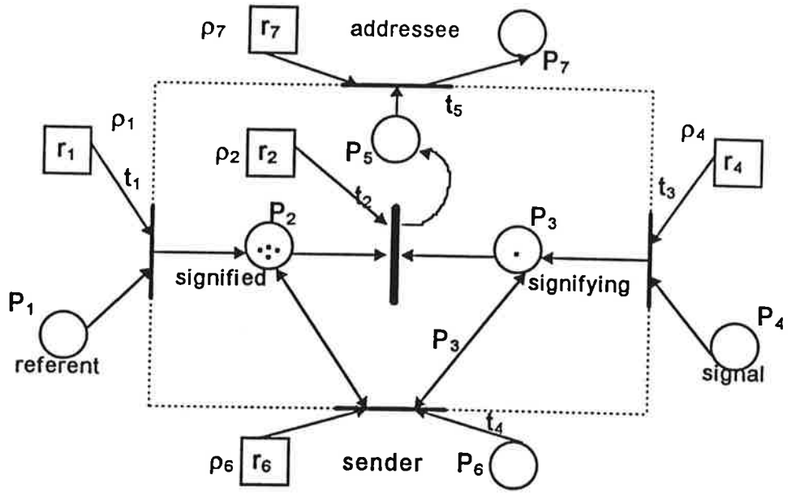


Fig. 2.

(We have presented the D_ε-network at the last but one stage of semiosis process, i.e. before the transition t_2 has come into act ion).

For this network the multitudes P, T, C and the function M have the form:

$P=\{P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, \rho_1, \rho_2, \rho_3, \rho_4, \rho_5, \rho_6, \rho_7\}$,

$T=\{t_1, t_2, t_3, t_4, t_5\}$, $C=\{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8\}$,

$M=\{m_1, m_2, m_3, m_4, m_5, m_6, m_7, m_8\}$.

References:

[1] Melnicov G. P. Systemology and language aspects of cybernetics. Moscow. Sovietskoe radio, 1978. 368 p.
[2] Automatic methods of text analysis and synthesis. /Piotrowski R. G., Bilan V. N., Borkun M. N. and others. Minsk: Vysheyshaya shkola, 1985. 222p.
[3] Mukhamedov S. A., Piotrowski R. G. Engineer Linguistics and system-statistic researches of Uzbek texts. Tashkent. FAI, 1986. 360 p.
[4] Nariniany A., Yakhno T. Production systems «Knowledge representation in man-machine and robot-technical systems/VINITI. Moscow. 1989 T. A., pp. 136-177.
[5] Piotrowski R. G. Engineer linguistics and language theory. Leningrad: Nauka, 1979, 122 p.
[6] Piterson G Petry network theory and systems modeling. Translated from English. Moscow. Mir, 1986. 264 p.
[7] F. de Saussure. Cours de linguistique generale. Translated from French. Moscow: Progress, 1977. 488 p.
[8] Frege G. Meaning and denotatum //Semiotica y informatica, 1977, ed. 8. Pp. 181-210.
[9] Chizhakowski V. A. Semiotic and communicative aspects of automatic processing of document titles. DSc Phil. Dissertation. Leningrad: Leningrad State University, 1988, 32 p.
[10] Shingariova E. A. Semiotic bases of linguistic informatics. Leningrad State Pedagogical University, 1987. 81 p.
[11] Shreider Yu. A. The logic of signs systems. Moscow. 1974 120 p.

Summary of the paper for QUALICO97

Type: Short Paper.

Title: Analysis of Japanese vocabulary by the theory of Synergetic Linguistics

Author: Haruko Sanada-Yogo

Address: 2-1-26-721, Kyodo, Setagaya-ku, Tokyo 156, JAPAN

E-mail: 19959014@gakushuin.ac.jp

Affiliation: Research Fellow of the Japan Society for the Promotion of Science - Gakushuin University

Summary:
This is a paper for the analysis of Japanese vocabulary, especially words which were created after the Meiji Restoration(1868) for new meanings imported from Europe, applying the theory of Synergetic Linguistics. The relationships between Word Length, Word Age, Number of meanings and Word Frequency are measured.

Project note.

Topic Area: Lexicology.

Semasiological and Word-formational Processes in Natural Language Lexical Evolution

Anatoliy A. Polikarpov

Lomonosov Moscow State University
Faculty of Philology, Russian Language Dept.
Laboratory for General and Computational Lexicology and Lexicography

e-mail: polikarp@philol.msu.ru

1. Basing on the assumption of basic semantic micro-processes dominating in the history of any word meaning is possible to predict [Polikarpov, 1988; 1990; 1991; 1993; 1994; 1995] and to test[Karapetjants, Obukhova, Polikarpov, 1988; Karimova, Polikarpov, 1988; Polikarpov, 1995; 1997; Polikarpov, Kurlov, 1994; Kolodjazhnaja, Polikarpov, 1992; 1994; Kapitan, 1994; Breiter, 1994; Breiter, Polikarpov,1997; Savchuk, 1997] some most probable directions for the derivative micro-processes (e.g., change of some semasiological and word-formational parameters of a word) and for macro-evolution of a lexical system on the whole, due to various, but correlated changes of its parameters.

2. Basic microproceses are:

(1) abstractivization of any meaning, loss of components by it in its individual history;

(2) relatively more abstract character of each successive meaning appearing in the history of a word as compared to a maternal meaning.

All other processes and features on the microlevel of a lexical system evolution (processes on the level of a word) are derivatives from mentioned basic ones: changes of the tempo of acquiring new meanings, different rate of losses for meanings of different semantic quality, change of word-formational ability of a word during its history, etc.

Processes on the macro-level of lexical system organization are the result of some specific integration of micro-processes under some boundary conditions specific for the system on the whole.

3. First of all, it is possible to construct a typical curve of the polysemy development of any word which integrates the information for the following basic processes:

- (1) some exaustion in time of the activity of any meaning in "generating" new meanings as a result of its "wearing out" (growth ofits abstractness) and making busy its associative valences in its own history as a result of accumulation of realised links while giving birth to new meanings ;

- (2) relatively lower initial level of any succesive meaning's activity in generating new meanings (as a result of its greater relative abstractness);

- (3) increase of the initial level of stability for any successive meaning (as compared to the corresponding preceding meaning) according to its correspondance to the relatively wider sense sphere;

- (4) slight growth of the initially obtained level of stability for a meaning while becoming more abstract in its history (but being usually exausted and deleted from

language life much faster than the mentioned kind of the stability can grow).

These four processes are present, but with different degree, in any word history .

4. Typical word polysemy development curve integrates step by step decrease of the tempo of acquiring new meanings up to zero and increase of longevity for each successive meaning up to some large (but not endless) term - in some asymmetric trajectory, with the exponentially-like retarding increase in the beginning, achieving some maximum and then having long history of decreasing polysemy, when addings of new meanings become lesser and lesser (stopping at some time), but losses grow (with retarding according to relatively greater stability of any successive meaning).

5. Processes of "relative" and "absolute" abstractivization of word meanings should lead to the greater probability of older words' meanings to enter synonymic and antonymic relations with meanings of other words in a vocabulary. It is explained by the fact that in the process of abstractivization specific, rare present components of meanings are more vulnerable, oftener are omitted. So, objective degree of similarity between all meanings of some degree of abstractivization should be greater than between meanings of a lower degree of abstractness. This fact leads meanings of ageing words to more successive finding of those counterparts which coincide with them in almost all components (and differ in some small proportion of them) than meanings of younger words, on the average, can do. Differential components in synonymic meanings typically can be neutralized in some specific (synonymic) conditions of their use, or, on the contrary, can be activated as symmetrically opposed (or, sometimes, opposed within a privative kind of an opposition).

Naturally, antonymic potential of a word also grows during its ageing, as a result of the same process of objective growing of similarity with meanings of other words and rise of probability for achieving clear and stable oppositions between meanings.

6. Ability of a word to "generate" new phraseologically bound meanings depends on two different factors, each of them being derived from the same basic factor of successive abstractivization of meanings, but changing in time their relevance for the considered process of idiomaticization in two opposite directions:

(1) the more abstract meanings of a word during its ageing, the less they are able to supply some components for fixing them within a new idiomatic meaning arising on the basis of intersecting meanings of several collocated words;

(2) the more abstract meanings of a word during its ageing, the greater variety of them are able to be used together with meanings of other words and, consecutively, greater chances for them to come across with some of others suit to them for building new idiomatic combinations. Combined action of the two factors developing in the opposite directions predetermines nonmonotonous character of the dependence of the development of word's idiomatic activity on its age.

7. It is natural to expect that the most probable (statistically dominant) direction of the categorial development within the nest of derivationally connected words will be the movement from some relatively concrete, objectively oriented categorial semantics of each word-base - towards its derivatives of more abstract and subjectively oriented categories (parts of speech). So, there should be a tendency to begin a word-formational tree mainly from nouns, to continue it with adjectives, verbs, adverbs, pronouns, etc., and to end it with words of pure syntactic quality like conjunctions and prepositions. This direction of categorial development most basically is predetermined by the

mentioned fact of the inescapable development of any word's integral lexical semantics during speech acts into the direction of the greater abstractness. More abstract lexical semantics seeks for the corresponding more abstract categorial form (which is more organic to it).

8. This succession of the categorial changes predetermines further that in any word-formational tree those words, which were produced on earlier steps of the process, should, on the average, be more derivationally active than words on further steps of it. It causes gradual decline of this process intencity which ends by the complete halt at some remote steps of it. It can be naturally explained by the narrower necessity in a language for words of greater categorial abstractness, than for words referentially more objective.

9. Among words of the same grammar category and of the same step of derivation the most active in "generating" new words should be the youngest ones. The older a word the less it is active in word-formational process. It can be explained by the greater amount of its de rivational potential waisted already on the previous stages of its productive existence.

10. Among words of the same age and of the same step of derivation, the most active in generating new words among other categories should be nouns. Adjectives and verbs should be weaker in this ability. The weakest should be semi-syntactic and pure syntactic words like prepositions and conjunctions.

11. Parallelly to the process of production of new words from some time there begins the process of words losses. Greater abstractness of the categorial semantics of later generated words of a nest also naturally presupposes greater, on the average, predisposition of them to longevity, stability. The most changing part of any language vocabulary are nouns, the most stable - pronouns and prepositions. Adjectives, verbs, adverbs and numerals are between them.

12. According to widening referential scope of words of greater categorial abstractness they should be characterized, on the average, by proportionally greater frequency of use than words of less abstract categorial semantics.

13. Initial stages of the word-formational process, as opposed to successive stages (analogously to the polysemy development within a word), should be characterized by the greater irregularity in semantic relations between bases and derivatives - in inheritance of meanings by derivatives from their bases. So, on initial stages of the process there should be observed oftener cases of pure "lexical" derivation (and oftener in its "mutation" variant than in its "modification" variant) and rarer - cases of pure categorial (so-called, "syntactic") derivation of words.

14. Polysemy of words on each next step of the derivational process should steadily decrease as a result of the additional categorial restrictions put by word-formative affixes of a new word on its lexical semantics inherited from its word-base.

15. The volume of any word-formational nest, as well as the number of meanings posessed by some word-forming affix, should develop with ageing of each of these units analogously to the development of word polysemy during aging of a word (according to the similarity of causes in both cases) - with the fast in the beginning, but gradually retarded growing of the number of elements in a unit (words in a nest or meanings in an affix), with arriving then at some maximum, when the growing intensity of the process of loosing elements (words in a nest or meanings of an affix) at last equals the intencity of

the process of acquiring new elements (words in a nest and meanings by an affix),and with prolonged, continiously retarded dicrease of nest's volume (or decrease of the number of meanings of an affix) in the course of its further existence.

16. Average length of morphemes belonging to words from nests of longer derivational history (and, correspondingly, of greater nest age) should be proportionally less than that of morphemes belonging to words from nests of shorter derivational history. This leads to existence of the so called Menzerath-Altmann's law. Natural explanation of this law for a word (as a specific construction) is connected with the categorial and age ordering of morphemes within a wordform.

17. Abstractivization of morphemes' meanings leads to the decline of clarity in boundaries between them and to the simplification of morpheme structure of a word, i.e., reinterpretation of a group of morphemes as some new morpheme. This is the most productive way for emerging new morphemes in a language.

18. Basing on these qualitative and quantitative assumptions and conclusons is possible also to predict, further, the most probable direction of the correlated development of the variety of the semasiological and word-formational features in relation to other language features - flectional, syntactic, phonological, etc. during some language's typological reconstruction (e.g., development from relative syntheticity to the more analytic structure or in the opposite direction).

19. Obtained data on some Slavic (Russian and Polish), Germanic (English and German), Romance (Latin, Italian, French, Spanish, Portuegese), Finno-Ugric (Finnish, Estonian and Hungarian), Turkic (Volga-Tartarian, Turkmen, Azerbaidjanian), Mongolian, Syno-Tibetian (Chinese), Vietnamese languages show clear correspondence between theoretically drawn and empirically investigated system regularities in the dependences between various language features.

References

BREITER M.A. [1994].
Length of Chinese Words in Relation to their other Parameters // Journal of Quantitative Linguistics. V.1, N3, 1994.

BREITER M.A., POLIKARPOV A.A. [1997].
Polysemy and Frequency of Word in Chinese: Experimental Study of System Dependences // IV International Conference on Languages of the Far East, South-East Asia, and Western Africa, September 17-20 1997, Institute of Asian and African Countries of Moscow University. M:, 1997.

KAPITAN M.E. [1994].
Influence of Various System Features of Romance Words on their Survival // Journal of Quantitative Linguistics. V.1, N3, 1994.

KARAPETJANTS A.M., OBUKHOVA N.I., POLIKARPOV A.A. [1988].
"Dictionary of Modern Chinese" (Peking, 1979) as a Sourse of Lexico-Typological Data) // Actual Problems in Studies of Chinese Language. Materials of the 4th All-Union Conference. Abstracts of Papers. - M., 1988.

KARIMOVA G.O., POLIKARPOV A.A. [1988].
Slovar' "Novye slova i znachenija " (1984) kak Istochnik Dannykh ob Evolutsii Jazyka ["New Words and Meanings" Dictionary (1984) as a Sourse of Data on Language

Evolution] // Applied Linguistics and Automatic Text Analysis. Papers from the All-Union Conference held 28.01-30.01.1988 at Tartu University).- Tartu: Tartu University Press, 1988.

KOLODJAZHNAJA L.I., POLIKARPOV A.A [1992].
Issledovaniie Sistemnykh Parametrov Leksiki na Osnove Komp'iuternoi Versii Sinonimicheskogo Slovarja (Study of Systemic Lexical Parameters Using a Computer Version of a Synonymic Dictionary) // Trudy Mashinnogo Fonda Russkogo Jazyka. Vol.II. - M.: Institute of Russian Language of Russian Academy of Sciences, 1992.

KOLODJAZHNAJA L.I., POLIKARPOV A.A.(1994).
Study of quantitative correlations between Stylistics, Grammar, and Polysemy of Words (On the basis of Ozhegov Dictionary). // Second International Conference on Quantitative Linguistics, Qualico'94, 20-24 September 1994. Moscow State University. Abstracts of papers. Moscow, 1994.

KUSTOVA G.I., POLIKARPOV A.A. [In press].
Novye znacheniya: tendentsiya k povysheniyu abstraktnosti (New meanings: Tendency to Growth of Abstractness) // System Studies in Linguistics - I. - M. (In press).

POLIKARPOV A.A. [1993].
A Model of the Word Life Cycle // Contributions to Quantitative Linguistics / Ed. by R. Koehler, B.B. Rieger. - Dordrecht: Kluwer, 1993.

POLIKARPOV A.A. [1994].
Zakonomernosti zhiznennogo tsikla slova i evolutsija jazyka. Statja 1. Modelirovanije osnovnykh sistemnykh sootnoshenij (The Regularities of Word Life Cycle and Language Evolution. Article 1. The Modelling of the Main System Correlations) // Russkij Filologicheskij Vestnik (Russian Phylological Bulletin), N 1, 1994. - Moscow, 1994.

POLIKARPOV A.A. [1995].
Zakonomernosti zhiznennogo tsikla slova i evolutsija jazyka. Statja 2. Teorija i eksperiment (The Regularities of Word Life Cycle and Language Evolution. Article 2. Theory and Experiment) // Russkij Filologicheskij Vestnik (Russian Philological Bulletin), N 1, 1995. - Moscow, 1995.

POLIKARPOV A.A. [1997].
Lexical Sybsystem of Natural Language System: Theoretical and Experimental Aspects of its Coming-to-Be and Evolutionary Study (in Russian, a manuscript). - Moscow, 1997.

POLIKARPOV A.A., KURLOV V.Ya., [1994].
Stylistics, Semantics, Grammar: Experience of System Correlations Analysis (On the Basis of Data from the Explanatory Dictionary) // Voprosy Jazykoznanija (Journal "Linguistic Problems"). N 1, 1994 (in Russian).

SAVCHUK L.O. [1997].
System Correlations between Vocabulary of the Society and that of an Individual. Ph.D. dissertation. - Moscow, Faculty of Philology of Moscow Lomonosov State University. - M., 1997.

ON MEASURING "TERMNESS":

A QUANTITATIVE APPROACH TO "TERM-NONTERM" CONTROVERSY

S.D.Shelov

Russia, Moscow 117 333, Vavilov str., 48, app. 335

e-mail: shelov@ippi.ac.msk.su
tel: (095) 137 58 83
fax: (095) 209 05 79

Committee for Scientific Terminology in Fundamental Research, Moscow, Russia

IT IS SUGGESTED THAT WE CAN MEASURE THE "TERMNESS" OF A LEXICAL ITEM ACCORDING TO THE NUMBER OF CONCEPTS NECESSARY TO IDENTIFY TERM'S MEANING. THE IDEA IS IMPLEMENTED THROUGH SPECIAL ANALYSIS OF TERM DEFINITIONS

TOPICAL PAPER
TOPIC AREA: 1;4

A few years ago I assumed that the nature of the term can be characterised in the following way (Shelov 1982; Shelov 1990):

1. it is a concept denoted by a lexical item (word or word combination) that makes this item a term,

2. termness of an item (= quality of being a term) is determined by all items necessary for the identification of its concept within the whole system of definitions (explanations) of these items, belonging to the field of knowledge under consideration,

3. the more information is required to identify a concept, denoted by a certain item, the greater its termness is.

Hence the concept of termness is postulated as purely relative, and some items are declared to be "more terms" and the others are declared to be "less terms". Roughly speaking, the greater is a conceptual addendum we are ready to accept in identifying a lexical item's meaning, the greater is the "termness" of this item.

Then immediately the question rises: provided with a good definition system of some lexical items, can we measure termness of any item, which is defined in this definition system?

Here I am going to present a positive answer to this question which, I believe, is of some theoretical and practical value for terminology, terminology data banks, knowledge engineering practice and so on. Measuring termness might also render a good service for any professional teacher as it helps in assessing difficulties a learner could face in his attempts to understand terminological items of any domain. Moreover, one also gets the opportunity to assess termness of any special text /domain (by summing up "termness" of all special items of this text/domain), the average termness of any special text/domain and so on.

I will assume here that there exists a consistent logically and linguistically irreproachable definition system for terms of a given domain. This assumption involves that any polysemy or synonymy of the expressions that defines the corresponding concept (definiens) is eliminated. Particularly, this means that every common word of the definiens expression has one and the same meaning, every syntactic relation sticks only to one and the same semantic relation, not a single linguistic meaning is expressed in different ways, etc.

If this holds true, I could consider as the amount of "information required to identify a concept, denoted by a certain item" just quantity of all concepts expressed in the definiens expression for this item, namely, quantity of all autonomous words used in the definiens expression. Thus we admit that any autonomous common word of the definiens expression contributes equally to the termness of the definiendum, i.e. the lexical item to be defined. According to the above stated, it contributes one and only one concept, so I will asses this concept contribution as 1.

Let's consider the definition system of computer science lexicon as this system is presented in the Webster new world compact dictionary of computer terms, compiled by Laura Darcy and Louise Boston (N.Y., 1983). Here I will put under consideration the following 17 lexical items: *arithmetic operation, bit, character, computer, data, digit, error, execution, fatal error, fixed-point notation, fixed-point operation, font, frame 3, instruction, program 1, radix point, storage* (figures "1" and "3" are to point out that we have picked up only the first and the third meanings of the corresponding items of the above mentioned dictionary). Note that we have the following definitions of the lexical units *arithmetic operation, digit, error, radix point*(slightly shortened and altered to be more explicit and consistent):

(Definiendum) Term	Definiens
arithmetic operation	The addition (1), subtraction (1), multiplication (1) or division (1) of numerical (1) quantities (1).
digit	Any of the symbols (1) representing (1) the positive (1) integers (1) in some numbering (1) system (1).
error	Any deviation (1) of a quantity (1) from the known (1), correct (1) value (1).
radix point	A period (1) that separates (1) the integer (1) portion (1) of a number (1) from the fraction (1) portion (1).

I have put figure "1" each time an autonomous word is encountered in the definiens expression making a concept contribution to the definiendum item as 1. Summing up and denoting "termness" of an item x as T(x), we have:

T *(arithmetic operation)* = 6, T *(digit)* = 6, T *(error)* = 5, T *(radix point)* = 7.

We also have the following definitions of the lexical units: *bit, character, computer, fixed-point notation* :

(Definiendum) Term	Definiens
bit	A *digit* (6) in the binary (1) number (1) system (1) represented (1) by 0 (1) or represented (1) by 1 (1).
character	An alphabetical (1) letter (1), *digit* (6) or special (1) symbol (1).
computer	An electronic (1) device (1) for performing (1) high (1)-speed (6) *arithmetic operations* (6) and logical (1) operations (1).
fixed-point notation	The representation (1) of a number (1) where the *radix point* (7) is assumed (1) to be (1) in a fixed (1) position (1).

Using the same notation, we get T *(computer)* = T *(arithmetic operation)* + 7 = 13, T *(character)* = T *(digit)* + 4 = 10, T *(bit)* = T *(digit)* + 7 = 13, T *(fixed-point notation)* = T *(radix point)* + 6 = 13.

Implementing the same routine we get the following results:

T *(data)* = T *(character)* + 6 = 16, T *(fixed-point operation)* = T *(arithmetic operation)* + T *(fixed-point notation)* + 2 = 21, T *(font)* = T *(character)* + 6 = 16, T *(frame 3)* = T *(bit)* + 3 = 16;

T *(storage)* = 2 T *(data)* + 11 = 43, T *(instruction)* = T (bit) + T *(character)* + T *(computer)* + T *(data)* + 8 = 60;

T *(program 1)* = T *(instruction)* + T *(computer)* + 9 = 82;

T *(execution)* = T *(instruction)* + T *(program 1)* + 4 = 146;

T *(fatal error)* = T *(error)* + T *(program 1)* + T *(execution)* + 1 = 234.

These somewhat fragmentry results completely match our intuition in accordance with which every lexical item participating in a definition of the other is of less termness than the latter. They also give a good picture of the "level structure" of terminology concept system: thus, the items *arithmetic operation, digit, error, radix point* are at the first concept level; the items *bit, character, computer, fixed-point notation* are at the second concept level and so on up to the item *fatal error* which takes its position at the seventh level. It is also worth drawing attention to the relation "to be defined through (the other term/terms)", since that is the terms already defined that contribute most of all to the termness of the items to be defined. This relation is of principle importance in terminology semantics apart of the problems discussed as it also plays a fundamental role in conceptual structuring of terminology, and particularly, in hierarchic genus-species relationships between terms (Shelov 1996). The results illustrated here give one more reason in support of this statement.

Bibliography

Shelov S.D. (1982) The Linguistic Nature of the Term //Automatic documentation and mathematical linguistics. - V.16. - N 5.

Shelov S.D. (1990): Terms, termability and knowledge //TKE^90: Terminology and Knowledge Engineering. - V.1. Frankfurt/M.

Shelov S.D. (1996) Concept structure of terminology and knowledge representation procedure //TKE^96: Terminology and Knowledge Engineering. Frankfurt/M.

On granularity in the interpretation of *around* in approximative lexical time indicators.

Paper submitted for the Third International Conference on Quantitative Linguistics (Qualico 1997), August 26-30, 1997, Helsinki (Finland)

Filip Devos, Patricia Maesfranckx and Guy De Tré
Department of Dutch linguistics and Computer Science Laboratory
University of Gent
Blandijnberg 2
B-9000 Gent
Belgium
Tel: +32/9/264.40.82.
Fax: +32/9/264.41.70.
E-mail: filip.devos@rug.ac.be
patricia.maesfranckx@rug.ac.be
guy.detre@rug.ac.be

Summary

The representation of approximative lexical time indicators (ALTI's) in natural language for building conceptual models, for integration in database systems or (other) AI-applications is made difficult by a number of factors. In this paper, ALTI's are discussed relative to two related aspects: (1) vagueness and (2) interpretation. As for (1), ALTI's are shown to be vague in degree. As for (2), granularity is considered to be the determining factor in the interpretation and representation of ALTI's.

Topical paper

Topic area

Time indicators, knowledge representation of time, vagueness, possibility theory, fuzzy set theory, documentation and information retrieval, databases.

1. Introduction

The representation of approximative lexical time indicators (ALTI's) in natural language for building conceptual models, for integration in database systems or (other) AI-applications is made difficult by a number of factors. In this paper, ALTI's are discussed relative to two related aspects: (1) vagueness and (2) interpretation. As for (1), ALTI's are shown to be vague in degree. As for (2), granularity is considered to be the determining factor in the interpretation and representation of ALTI's.

The first part of the paper analyses the vagueness of lexical time indicators (LTI's) in general (§2), both as far as their status (§2.1) and as far as their interpretation is concerned (§2.2). In the second part of the paper (§3), approximative lexical time indicators (ALTI's) are singled out for an analysis along three lines: (1) the importance of modelling ALTI's (§3.1), (2) the interpretation of ALTI's (§3.2) and (3) specific factors determining the interpretation of ALTI's (§3.3).[1]

2. Lexical time indicators (LTI's)

'Time' is an extremely complex notion, as in natural language different time conceptions and divisions are reflected: (1) physical or natural time (e.g. *day* as the time it takes for the earth to turn around its axis), (2) artificial or calendar time (e.g. *century* as a period of 100 years), and (3) experiential or psychological time (e.g.. *evening* as the period between work and sleep). The categorization of time is mostly determined by convention and on the basis of natural regularities (Devos et al., 1994). In language, time is reflected in different ways: in tenses, aspect, lexical items, numerical elements or a combination of these. This paper deals with lexical time indicators which may contain numerical elements.

2.1. LTI's and vagueness

LTI's indicate either time position (e.g. *today, shortly before 6 p.m.*), frequency (e.g. *3 times a year, often*) or duration (e.g. *the whole day, about 3 hours*). Apart from this categorization, LTI's can be subdivided according to the following parameters:

(1) relational-situational:
Relational LTI's refer to a relation with a time point or interval and this relation is an anterior (e.g. *shortly before 6 p.m.*), a posterior (e.g. *some years after the war*) or an approximative one (e.g. *around 10 a.m.*). Situational LTI's point to a time fact itself (e.g. *in May, at 10 a.m., last year*).

(2) bound-unbound:
Unbound LTI's do not refer to past, present or future (e.g. *at two o'clock, in May*). Bound LTI's, on the other hand, do refer to past, present or future (e.g. *at two o'clock yesterday, in May 1944*).

Some of these expressions contain "vague" information: the extension of expressions such as *shortly before 6 p.m.* may be said to be fuzzy, as one may wonder wether *5.40 p.m.* still falls within the extension of this time indicator. Vagueness should be distinguished from other forms of lexical polyvalence, such as ambiguity and generality, with which it is often confused. **Semantic vagueness** refers to an intrinsic uncertainty as to the application of a word to a denotatum (Devos, 1995). With **ambiguity** the uncertainty is not intrinsic, as it is situated only on the side of the hearer. If a speaker says: 'I'll call you *at 9 o'clock*' and it is not obvious from the context whether a.m. or p.m. is meant, the speaker has the choice between a limited range of possible interpretations. The speaker however, knows exactly which one is meant. This is not the case for vague expressions. **General** information is found especially in situational and unbound expressions, when they refer to an interval. In 'My birthday is in May' the information is unspecified or underdetermined, though the boundaries of the period are fixed (i.e. *between 1st and 31st May*), as opposed to vague expressions.

Time indicators indeed often show some **vagueness in degree**, as opposed to vagueness in criteria (Devos, 1995). The first kind of vagueness resides in the fact that one and only one well-determined criterion is being scaled (e.g. the criterion "age" in *an old man*). **Vagueness in criteria**, on the other hand, can be found in expressions like *a big house*: most often different criteria are called upon in naming a building *a big house*. Hence, this kind of vagueness is multidimensional. Many lexemes are vague in both senses. Semantic vagueness can be found especially in the following subclasses of LTI's: (1) lexical, non-numerical indicators of frequency (e.g. *often, seldom*); (2) approximative time indicators (e.g. *around 6 p.m., around 1972*) and (3) indicators of half closed (or half open) intervals, i.e. indicators of posterior and anterior relations (e.g. *shortly before 6 p.m., some time after the holidays*). These are the three types which we have investigated, as described in §2.2 below.

As time is a one-dimensional fact, vagueness in degree is involved. Moreover, time can be expressed numerically (which makes time objectifiable). All this should facilitate a formal representation of vague LTI's by means of fuzzy set theory. In Devos et al. (1994) and Van Gyseghem et al. (1994) an analysis is given of a formal way of representing vague LTI's by means of fuzzy set theory, probability theory and fuzzy logic. They outline different models of representing vague time intervals by means of fuzzy set theory. It is argued that this differentiation is needed if the (combined) data obtained through inquiries are to be modelled into a single fuzzy time interval that is suited as the representation of a linguistic term[2].

2.2. The interpretation of three sorts of vague LTI's: inquiry

In order to create an experimental basis for the representation of the semantics of vague LTI's a survey inquiry was carried out[3]. Informants were asked what the underlined time indicators referred to in 16 sentences. No predetermined answer possibilities were given. The vague LTI's which appeared in the sentences were of the three types mentioned in §2.1: (1) lexical, non-numerical indicators of frequency (*often, now and then, seldom*); (2) approximative time indicators (*around 8 p.m. last night, around the turn of the century, around Easter, around 5.10 p.m. and around lunch-time*) and (3) indicators of half closed (or half open) intervals, i.e. indicators of posterior and anterior relations (*after the Second World War, at the end of next week, early this morning, until deep in the 19th cetury, the last few weeks, of short duration, in the near future, shortly before 6 p.m*). The results of this inquiry shed a light on some (cognitive) principles which determine the interpretation of vague LTI'S by average language users:

a) symmetrical intervals for ALTI's
Symmetry seems to be very important in the interpretation of the second type mentioned above (ALTI's). An overwhelming majority of the answers consisted of symmetrical intervals around the reference point given. The symmetry was only broken if for instance round numbers are used.

b) round numbers
Round numbers function as cognitive reference points in the numerical system (Channell, 1994). In our inquiry it is quite obvious that approximations are mostly given in terms of round numbers. An approximation of 10 years is more likely to appear than an approximation of 9 or 11 years. Roundness can also explain the asymmetry in some answers (e.g. for *around Easter* = 11th April, there were answers like: 1-30 April, 1-20 April and 1-15 April). This is often connected with the avoidance of granularity shifts, as outlined in §3.3.2.

c) experiential factors
Whereas the values given for ALTI's are quite uniform across informants, there is a much larger variation in the answers given for the first and the third type of time indicators included in the investigation. This can be explained by experiential factors. For instance, in *at the end of next week* 'week' was interpreted by some informants as ending on Friday (school/working week), by others as ending on Sunday (normal week). For lexical frequency indicators (*often, seldom, ...*) experiential factors seem to be extremely important. The values given for *now and then* in the sentence '*He only drinks alcohol now and then.*' range from 0-2 times a month till 8-12 times a month, most probably due to the informants' own experience with alcohol.

The inquiry has shown that the semantics of ALTI's will be more easy to formalize than that of the other types, as there is more agreement among language users about their meaning, which is less dependent on experiential factors. Therefore a new inquiry was set up, concentrating on ALTI's.

3. Approximative lexical time indicators (ALTI's)

In the second part of this paper we focus on ALTI's, indicators of time that render a point of time or a time interval approximatively (e.g. *around six o'clock, around 1974, around the turn of the century).* They consist of "approximators" and "approximata". Approximators are always lexical items (e.g. *around*), approximata are either lexical (e.g. around *noon*) or numerical (e.g. around *8h30*). In ALTI's, the gradual vagueness is to be found in the modifying expression (e.g. *around* 6 p.m.) or in both the modifying and the modified expression (e.g. *around noon*). ALTI's refer to an (symmetrical) interval stretching between two vague or fuzzy limits. We will focus on factors determining the interpretation of *around* in ALTI's having numerical elements as approximata (§3.2.). Not only are lexical items as approximata (e.g. *evening*) often vague in criteria, but as a rule they are also more easily experientially determined (e.g. *evening* as 'period between day and night' or 'period between work and sleep' or 'period of rest after work').

3.2. Theoretical model

In most methods of representing time the traditional view on categorization is reflected: time indicators are sharply delimited and their application is either true or false. For instance, next to the usual symbols from ordinary logic, tense logic (Prior, 1967) uses special symbols, i.e. the time operators H(abitually), G(enerally), P(ast) and F(uture), for rendering time indicators, but time is traditionally conceived of. Moreover, tense logic focuses on tenses, not on lexical time indicators, and on modal logic. Tenses, however, are only one, though substantial and (proto)typical means of indicating time. Not only tense, which has received most attention in the literature, but also temporal prepositions, adverbs and open class lexical items, especially nouns, as well as word order, amongst others, determine temporal reference. Our analysis of ALTI's shows the traditional view to be incorrect.

Instead, fuzzy set theory (FST)(Zadeh, 1965, 1974) can be used as a model of representing vague lexical items. A central claim of FST is that category membership can be a matter of degree. Fuzzy (closed) intervals can be used to model vaguely expressed periods of time. A fuzzy time interval is an immediate fuzzy extension of the crisp notion of time interval: whereas a point of time x either does or does not belong to a crisp time interval, it can "belong" to the fuzzy time interval, modelled by the membership function m, with a degree m(x).

Vague LTI's, especially ALTI's, are pre-eminently analysable in FST-terms, because: (1) 'time' is an unidimensional notion; i.e. time indicators have but one base variable, and are thus vague in degree; (2) time data are clearly expressable in numerical data and time can be objectified, and (3) the predominant symmetrical values given in the interpretation of ALTI's facilitate the formalization. For instance, *around 18h* can be given the following values (mentioned between square brackets) for 17h30 [0], 17h40 [0.4], 17h50 [1], 18h10 [1], 18h20 [0.4] and 18h30 [0], meaning that the time points with value 1 fully belong to the extension of *around 18h* and the ones with value 0 do not belong to it. The results from the inquiry can be converted into intervals represented by functions in block model or by bell-shaped functions.

3.2. Inquiry on ALTI's

To determine the interpretation of the fuzzy intervals ALTI's refer to, an inquiry was carried out in which informants were asked to intuitively indicate sharp (crisp) and closed time intervals for a range of expressions with a granularity differing in level and size (cf. §3.3.1): *around April 28th, around 20h10, around 2070, around 10h57, around 4000 B.C., around September 1993, around 19h30, around 350 A.D., around 1974, around (14h) 10min 05sec, around February 3rd, around 1979, around 14h15, around 2000, around 18h, around March 15th, around 1670, around (12h) 12min 17sec, around 18h22* and *around 2500*.

A methodological shortcoming of the first inquiry (cf. §2.2) was that it could not be derived if informants cognitively represent the meaning of ALTI's as a fuzzy or a crisp interval, as they were asked to give a crisp interval. In the second inquiry this problem was solved by asking them to indicate two intervals. For instance, for *around 1974* the informants were asked to indicate the interval the ALTI definitely does refer to (Y = yes; e.g. *1972-1976*) as well as the interval it definitely does not refer to (N = no: e.g. ←*1970-1978*→). In this way it could be derived whether the Y and the N values are adjacent or if there is a zone in between. This method gave a maximum and a minimum value for each ALTI. For instance, for *around 1974* "max Y" was *1971 -1977*; "min Y" was *1973 -1975*; "max N" was ←*1970-1979*→ and "min N" was ←*1972-1976*→. From these data a uniform nuclear interval could be derived for each ALTI. In each case the intervals taken had to represent 20% of the answers minimally. In 75% a Y-interval was identical to a N-interval. In 60% the value for the Y max-interval was identical to the value for the N min-interval. However, in only 25% of the individual answers the Y-interval was identical to the N-interval (especially in the ALTI's *around (14h) 10min 05sec*, *around 14h15* and *around 2500*, and remarkably not in, for instance, *around 18h22*). This implies that people only sporadically and unsystematically give crisp or sharp intervals for ALTI's. In most cases they do not define a sharp border between Y and N, reflected in the fact that they leave some space in between. It was hoped that some aspects at least of the interpretation of ALTI's would be determinable.

3.3. The interpretation of ALTI's

3.3.1. Granularity

The results of the inquiry showed granularity to be one of the determining factors in the interpretation of ALTI's. Granularity refers to the (abstract) time levels people use. It constitutes a rather precise hierarchical system of subordinate and superordinate categories in which different shifts may occur (e.g. *second* → *minute* → *hour* → *day* →...). The cycli form different levels, which are not always and not all relevant for the interpretation of expressions or sentences containing ALTI's. For instance, the age of an infant is often expressed in terms of months (e.g. Our daughter is *14 months* old now), though the age of older children and adults is referred to by years only.[4]

In their system for the automatic deduction of temporal information, Maiocchi et al. (1992) use five levels of granularity, with year as an 'absolute' datum and month, day, hour and minute as cyclical data:

YEAR (year XXXX - year XXXX) e.g. *around 1979*
MONTH (month 01 - month 11) e.g. *around September 1979*
DAY (day 01 - day 27-30) e.g. *around February 3rd*
HOUR (hour 00 - hour 23) e.g. *around 6 p.m.*
MINUTE (min 00 - min 59) e.g. *around 6.10 p.m.*

Concerning this five-level granularity, the following questions arise:

(1) As noted above, more levels can be distinguished in principle, for instance: *second* → *minute* → *hour* → *day* → *week* → *month* → *season/trimester/semester* → *year* → *decade* → *century* → *millennium*. This list is not exhaustive (e.g. *picosecond* and language-specific notions as English *a fortnight* or French *une quinzaine*). How many and which levels are to be taken?

(2) Undoubtedly some levels are cognitively more salient than others. Periods of a second, for instance, are not that important in everyday life, though periods of an hour are: our plan for the day is mainly based on it. It may well be that periods of half an hour, a quarter, 5 minutes or 10 minutes are equally important, though concepts like "period of 5 minutes" or "period of 10 minutes" are not lexicalized in language. This could undermine the above-mentioned reduction as for its cognitive basis. Indeed, from the inquiry the hypothesis can be derived that ALTI's can be given the following intervals for each level of granularity[5]:

second: approximation of 5 seconds (*around (14h) 10min 05sec, around (12h) 12min 17sec*)
minute: approximation of 5 minutes (*around 18h22*)
hour: - full hour and half hour: approximation of 15 minutes (*around 19h30, around 18h*)
- hour + nx5 min (= multiple of 5 min): approximation of 5 or 10 minutes (*around 20h10, around 14h15*)
- hour + nx1 min: see minute
day: approximation of 7 days (week) (*around February 3rd, around March 15th*)
month: approximation of a fortnight (*around September 1993*)
year: - *around 4000 B.C.*: approximation of 50 to 500 years
- *around 2500*: approximation of 50 years
- *around 350 A.D., around 1670, around 2070*: approximation of 10 years (decade)
- *around 1974, around 1979*: approximation of 2 years
- *around 2000*: approximation of 2 to 5 years

This implies that other levels than those mentioned by Maiocchi et al. (1992) are important in the conceptual interpretation of ALTI's.

(3) Some inclusion relations have to be normalized, for instance in *month* (either 28, 29, 30 or 31 *days*) or *year* (either 365 or 366 *days*). Some default values will have to be postulated anyway.

3.3.2. Shifts of granularity

It is not enough to simply postulate a granularity scheme, one should also look at the functionality of such a scheme, and see if, for instance, the existence of one level of granularity acts as a brake on the given value for a sublevel of granularity. In other words, does the place of the approximatum relative to a higher or lower level in the hierarchy have any influence on the interpretation of the ALTI? And if so, when does a level shift occur, i.e. when does the interpretation shift to the superordinate of the approximatum? An answer to these questions has significant implications for the symmetrical character of ALTI-intervals. The values given for the intervals are almost without exception symmetrical. For instance, for *around 18h*, in which the approximatum refers to a full hour, a value *between 17h45 and 18h45* can be given, but does this hold for *around 18h20* (= *18h05 - 18h35*) or *around 18h18* (= *18h03 - 18h33*)? Does the interpretation exceed the half hour level in these cases? There is, for instance, no shift of level in *around April 28th, around February 3rd, around 10h57 or around (14h) 10min 05sec* (with respectively *May 1st, February 1st, 11h and (14h) 10min* as endpoints). This implies that some intervals may be asymmetrical: *around 10h57* is

predominantly valued as *10h55 - 11h*, *around 1979* as *1977-1980* and *around 18h22* as *18h15-18h30*. In *around 4000 B.C.*, however, the shift necessarily does take place (Y min *4100 - 3900*, Y max *4500 - 3500*, N min ← *4500 - 3500* → and N max ← *5000 - 3000* →), as it is a round number, situated on the border of two granularity levels.

From this we must conclude that there is a correlation of factors. Not only granularity itself but also the roundness of the reference points and the position within this granularity determine the interpretation. In a previous inquiry, the ALTI *around April 11th* got an asymmetrical interval, due to the fact that it falls almost in the middle of the month. Of equal importance as the predicate appearing in the proposition is the distance between speech time and reference time.

3.3.3. Speech time and reference time

Reichenbach (1947) subdivides linguistic (tense-related) time into: speech time (ST), reference time (RT) and event time (ET)[6]. ST is the time at which an expression or sentence is uttered, RT the moment which is referred to and ET the time at which what is reported on takes place. Important for the interpretation of ALTI's is the distance between RT and ST: in general, a small distance (e.g. RT = *around 2000*; ST = *1997*), diminishes the value of the interval, while a big distance (e.g. RT = *around 2500*; ST = *1997*) enlarges this value. However, this rule does not seem to apply (to the same extent) to smaller granularities, like second, minute or hour (e.g. ET = *around 19h30*; ST = *18h*). In some cases ST acts as the limit of the interval, as can be seen in *around 2000*, which, in 1997 will be valued as *1997 - 2005*.

3.3.4. The size of the approximatum

A factor correlating with the previously mentioned distance between RT and ST, is the size of the approximatum: the larger level the approximatum refers to, the larger the interval is valued. Undoubtedly, *around 4000 B.C.* has a bigger interval than *around 3850 B.C.* or *around 350 A.D.* From the inquiry it is also clear that the smaller the level of granularity of the approximatum, the more agreement there is on the given intervals. There is more consistency in the values given for *second, minute* or *hour* than in the values given for *day* or *year*.

4. Conclusion

Our analysis has shown that ALTI's are vague in degree. In the interpretation of ALTI's a number of factors, the most important of which is granularity, seem to correlate. In theory, a fuzzy set-theoretic approach to ALTI's has a surplus value to other approaches in that it cognitively more adequately deals with vague expressions. FST can model ALTI's more adequately than traditional representations, though some problems remain. This modelling is important for the representation and handling of LTI and for time reasoning.

Typically, however, is the lack of cognitive (linguistic) evidence for this modelling. Dubois and Prade (1989), for instance, develop a system for representing vague time knowledge using FST, but their analysis presupposes an adequate linguistic description, which in turn presupposes the possibility of such a uniform and complete description. Only in a summary they point out the personal and contextual dependence of temporal knowledge.

This paper argues for a more global (cognitive) approach to vague approximative time indicators and for an adequate linguistic description. This linguistic description should be based on findings about how human cognition handles and organizes time data. It should, amongst others, take into account the three time conceptions mentioned in §2, and the various interrelated factors determining the interpretation of ALTI's.

NOTES

[1] In this paper we report on results of a interdisciplinary research project, financed by the Fund for Scientific Research (Flanders), carried out at the University of Gent: 'Fuzzy Temporal Databases, approached interdisciplinary from linguistics and from computer science.' (Project number 3.G.0091.96).
[2] Some major objections towards fuzzy set theory and the relation between this theory and prototype theory are discussed at length in Devos (1995).
[3] The inquiries were held amongst fifty students of Germanic philology and twenty students of Informatics at the University of Gent (aged 18-22), in December 1992.

[4] The same holds for spatial expressions, which may also show different levels of 'granularity'. For instance, a question such as 'Where are you working at present?' may, depending on the context and the situation, yield different answers, like 'in Europe', 'in Belgium', 'in Gent', 'at the University of Gent' or 'in the Department of Dutch linguistics'.
[5] Mind that these prototypical intervals can be changed by the factors described in §3.3.2.
[6] For comments on this division, see Klein, 1994.

REFERENCES

Channell, J. (1994), *Vague Language*. Oxford University Press.

Devos, F. (1995), 'Still Fuzzy after All These Years. A Linguistic Evaluation of the Fuzzy Set Approach to Semantic Vagueness.' In: Quaderni di Semantica, 16-1, pp. 47-82.

Devos, F., N. Van Gyseghem, R. Vandenberghe and R. De Caluwe (1994), 'Modelling Vague Lexical Time Expressions by Means of Fuzzy Set Theory.' In: Journal of Quantitative Linguistics, 1-3, pp. 189-194.

Devos, F. (1996), 'Semantic vagueness and lexical polyvalence' (submitted).

Dubois, D. and H. Prade (1980), Fuzzy Sets and Systems: Theory and Applications. New York: Academic Press.

Dubois, D. and H. Prade (1989), 'Processing Fuzzy Temporal Knowledge.' In: *IEEE Transactions on Systems, Man, and Cybernetics*, 19-4, pp. 729-744.

Klein, W. (1994), *Time in Language*. London: Routledge.

Maiocchi, R., B. Pernici and F. Barbic (1992), 'Automatic Deduction of Temporal Information.' In: *ACM Transactions on Database Systems*, 17-4, pp. 647-688.

Prior, A.N. (1967). *Past, present, and future*, Oxford.

Reichenbach, H. (1947). *Elements of Symbolic Logic*, New York.

Van Gyseghem, N., R. Vandenberghe, F. Devos and R. De Caluwe (1994), 'Fuzzy Time Expressions in Natural Language Queries.' In: *Proceedings of FQAS '94. Workshop on Flexible Query-Answering Systems, November 14-15, 1994, Roskilde University, Denmark*, n.p.

Zadeh, L. (1965), 'Fuzzy sets.' In: *Information and Control*, 8, pp. 338-353.

Zadeh, L. (1974), 'The concept of a linguistic variable and its application to approximate reasoning.' In: *Information Sciences*, 8, pp. 199-249, pp. 301-357, and 9, pp. 43-80.

L. Komarnytska

FACTORS DETERMINING PHONETIC MOTIVATION OF THE WORDS : AN EXPERIMENT IN PHONETIC SYMBOLISM.

A series of questions connected with the investigation of sound-sense correlation is known in linguistics as the problem of phonetic symbolism and phonetic motivation.

Traditionally, phonetic motivation (PM) is defined as a certain relationship between the sound form of the word and its meaning. Thus, 3 types of PM are differentiated :

1) positive PM (the sound form of the word is appropriate to its meaning);

2) negative PM (the sound form of the word contradicts its meaning);

3) zero PM (the relations between the sound form of the word and its meaning are neutral).

For the most part researches into the problem of PM were limited because the experimentators selected for the analysis only words with vividly revealed "expressive" component of meaning.

The objective of the present study was to obtain "quantitative measures" of PM of various lexical units and to examine some previously unexplored factors which might influence the degree of sound-meaning linkages.

Methods and Procedure

Stimulus Material

The list employed in the experiment contained 300 words chosen from the New English-Russian Dictionary (2) and 300 words taken from Thorndike-Lorge word count (4). As the selected words belonged mostly to the "neutral vocabulary", 100 additional words characteristic of special literary-bookish vocabulary (archaic,

historical words, terms, etc.) and non-standard vocabulary (dialectal words, jargonisms, vulgarisms, etc.) were chosen.

We included into the list lexical units which:

1) belonged to one of the three parts of speech (nouns, verbs, adjectives);

2) did not contain "meaningful" prefixes and suffixes (un-, dis-, etc.)

3) whose sound form did not resemble the sound form of the Russian or Ukrainian counterpart.

The order of presentation of 700 words was determined on a random basis.

Subjects and Questionnaires

70 undergraduates of Chernivtsi University (Faculty of Foreign Languages), majoring in English served as subjects of this study.

The written and oral instructions directed them to rate the stimulus word on a 5-point scale as to "fittingness" of the sound form of the word to its meaning.

"5" indicated the closest sound-sense linkages;

"4","2" - decreasing appropriateness;

"3" - neutrality;

"1" - complete lack of appropriateness of the sound form of the word to its meaning.

The questionnaires included:

1) an English word;

2) its transcription which guaranteed the adequate "sound image" of the word;

3) the Russian counterpart presented by 2 or 3 translation equivalents.

This technique of presentation the Russian counterpart (the first dictionary meaning) was aimed at elimination a possible tendency for subjects to seek for structural similarity between English and Russian sound forms.

Results and Discussion

The ratings of each word were summed and the summed scores for the words were averaged.

The data were processed with the help of the statistic criteria X^2 and the coefficient of concordance.

Only the results significant at the 0.01 level were taken into consideration.

The following factors which might influence the degree of PM of the words were analysed: semantic status of the word, its stylistic register, grammar category and frequency of usage.

The stylistic register and grammar category of the words were identified by the corresponding dictionary indications (2). The frequency class of the word was defined according to Thorndike-Lorge word-count (4).

The semantic category of each word was determined in Experiment II. The subjects were asked to classify 700 words into 14 "meaning" categories:

1) sound; 2) taste; 3) smell; 4) tactual experience; 5) light; 6) movement; 7) size, form; 8) colour; 9) intellectual and mental activity; 10) appearance; 11) features of character; 12) emotions; 13) areas of space; 14) others.

The statistic treatment of the data has given the following results:

1. The functioning of phonetic symbolism in the English language is for most part restricted to the words which denote sound, movement, smell, light, tactual experience, i.e. those concepts which present elements of the sensory continuum. The least phonetically motivated lexical units are the words denoting concrete and abstract notions. These results appeared to confirm the data obtained by J.M.Peterfalvi (for the French language) (3) and by J.Kurcz (for Hungarian, Chinese, Japanese and Swahili) (1). Thus, the experiment justifies the universal character of the phenomenon of phonetic symbolism.

2. The degree of PM depends on the stylistic layer of the word. Words which belong to non-standard vocabulary revealed closer sound-sense relationship than words characteristic of the special literary bookish vocabulary.
3. High-frequency words are characterized by stronger sound-meaning correlation than low-frequency words.
4. The relationship between the sound form of the word and its meaning is closer in verbs and adjectives than in nouns.

Summary

700 English words were measured by 70 subjects on a 5-point scale with the aim to define the appropriateness of the sound form of the word to its meaning.

The results of the experiment indicate that the degree of sound-sense correlation associates with several factors: semantic status of the word, its stylistic register, grammatical category , frequency of usage.

BIBLIOGRAPHY

1. Kurcz I. "Cultural and Linguistic Determinants of Phonetic Symbolism." International Journal of Psycholinguistics. 4.1 (1977) : 5- 11.

2. New English-Russian Dictionary. 2nd ed. 2 vols. Moscow: Russky Yazyk, 1979.

3. Peterfalvi J.M. Recherches Experimentales sur le Symbolisme Phonetique. Paris: Centre de la Recherche Scientifique, 1970.

4. Thorndike E., Lorge I. The Teacher's Word Book of 30 000 Words. New York: Teachers College, Bureau of Publications, Columbia University, 1944.

On the Nature of Koehler's Effect

Yu.K. Krylov

Moscow Lomonosov State University
Faculty of Philology
Laboratory for General and Computational Lexicology and Lexicography
e-mail: polikarp@philol.msu.ru

In the monograph by R.Koehler [1] there was described a very important effect of the words' average length oscillation as a function of their frequency of use (according to the data from a representative frequency dictionary). Use of the approach suggested by R.Koehler to the analysis of data (i.e., calculation of gliding averages not only for the function, but also for the argument) allowed us to establish that oscillation of the conditioned averages are observed also in other distributions. For instance, they take place in the case of the average words' polysemy dependence on frequency of words' use. Also it is observed for regression line of bimodal length distributions for units of different levels of coherent text organization. In this context, further as the Koehler's effect we will mean the polymodal character of the dependence of the conditioned average in case of two-dimencional distributions of any linguistic nature.

The main problem arising in connection with this effect study consists in search of answer to the question about the mechanizm of arising of observed nonmonotonous dependences. As it is known, smoothing of some time series determines change of its structure and, possibly, leads to the emergence of "low frequency" oscillations even in the case of pure random series - so called Slupski-Jule effect. Correspondingly, there is a problem of separation of the oscillations, contained in the empirical data and the oscillations generated by the calculation process, which can be resolved by use of different algorithms of averaging. That is why while revealing empirical dependences there were used not only various algorithms of gliding means calculations, but also smoothing with the help of geometric mean (averaging of logarithms for basic magnitudes), and even were used some robust algorithms not having some typical scale for a window of averaging.

Considering of the dependence of lexemic length L on frequency of their use F for vocabularies of M. Lermontov and V. Shukshin showed that under the condition of norming of the frequency spectrum by the frequency of the first rank word, position and width of the main maximum in these vocabularies coincide with the analogic characteristics in R.Koehler's studies [1]. Use of frequency dictionaries of textual conglomerates permitted us to establish that amplitude of observed oscillations decreases according to the increase of variety of texts involved in the textual conglomerate and according to the decrease of length of textual fragments taken for compiling of some frequency dictionary.

Analysis of the trend L = L(F) in case of novels by F.Dostojevski ("Demons"), M.Bulgakov ("Master and Margaret"), V.Belov ("Customary Case"), I.Turgenev ("Asia") and other whole texts allowed to find that the trend can be satisfactory described by the logarithmic dependence $L = a\ln F + b$. In this case coefficient b significantly increases as a function of the considered text length. The latter fact gives evidence for the statement that "satiation" of a text by lengthier words appears not locally in the vocabulary area of low frequency words, but more or less steadily along the whole frequency range.

Smoothing of remainders obtained by way of subtraction of the trend from the corresponding empirical dependences revealed presence of some stable maximums, appearing in the result of processing of some separate large parts of fiction prosaic texts, and the correlation of these maximums with oscillations observed in case of corresponding processing of the whole texts. However, most brightly the Koehler's

effect has manifested itself in case of considering the dependence of an average word length on the length of the whole text. Using data on the length (estimated both by the number of occurences of graphemes and runnig words) of more than one thousand short stories, stories, short novels and novels by various authors we managed not only to reveal presence of numerous maximums, but to connect extremums of observed curve with the typical scale for units (sentences, paragraphs, episodes, chapters) of different levels of the whole text organization.

All present permits to form a hypothesis, that the main complex of causes leading to Koehler's effect is connected with the hierarchy and interlevel interaction between structural elements in the whole texts. At the same time, just the effect of Slupski-Jule (however appearing on the deep level of text generation - as some consequence of averaging parameters for units of some lower level within some whole bigger fragment) leads to the rise of interdependent oscillations for size of classes of elements from different mesoscopic levels of text organization. This, eventually reveals itself in the Koehler's effect while considering corresponding distributions. Meanwhile, the major goal of the further research consists in necessity of closer relating of the locations of observed oscillations with the typical scales of different units obtained through the whole text specific segmenting.

References

[1] Koehler R. Zur Linguistischen Synergetik: Struktur und Dynamik der Lexik. - Bochum: Brockmeyer, 1986.

Three Laws of Fiction Prosaic Texts Organization

Yu. K. Krylov

Moscow Lomonosov State University
Faculty of Philology
Laboratory for General and Computational Lexicology and Lexicography
e-mail: polikarp@philol.msu.ru

There was undertaken a research of lengths of marinons, some integral units obtained by segmenting of comleted (whole) fiction prosaic texts (short stories, stories, short novels, novels). As natural marinons of different mesoscopic levels of text organization there were considered syllables, phonetic words, syntagms, sentences, paragraphs, subchapters, chapters, groups of chapters of a text, etc. Soft tools used in this case permitted to analyse statistical regularities for the text organization not only for the usually used grapheme display of it, but also for certain phonetic transcription of it. As the main rythmical sense-organizing unit in this case was taken a phonetic word defined as follows:

(1) it is any full (not syntactic) word presented in it transcribed form, or
(2) it is a combination of some full word with some syntactic word attached to it.

In both cases there should be one and only one phonetic accent within a unit.

Overall corpus of considered Russian fiction prosaic texts contains more than two million of graphic running words.

Undertaken research has lead us to revealing of three main regularities:

1) to the law of optimum relation between number of elements within each submicroscopic level;

2) to the law if free (unconditional) combination of phonetic words in a text;

3) to the law of fractal likeness for the distributions of lengthes of marinons excerpted from different levels of the whole text organization.

In accordance with the first law an optimal correlation between the overall number of phonemes and spaces (in a text segmented with the help of phonetic words) is characterized by two universal linguistic constants:

$$C_1 = (5 - \mathrm{sqrt}(5))/10 = 0.2764 \text{ and } C_2 = (5 + \mathrm{sqrt}(5))/2 = 3.618 \qquad (1)$$

The first of them determines the correspondence between the number of spaces in a text and the number of rithmons (number of vocal phonemes together with the number of spaces).

The second law determines correspondence between the number of consonant phonemes and the number of spaces (phonetic words). In this case C_1 and C_2 appear to be connected by the mutually inverse correspondence:

$$C_1 * C_2 = 1,$$

which reflects equality of numbers for consonant phonemes and rhythmons in an ideal case.

Theoretical correspondences (1) were obtained as a result of solving some optimizational problem. Their testing by looking through volumous empirical data showed that they follow the law with a precision characterized by the coincidence of the third significant figure.

Investigation of features of coupling combinations of phonetic words permitted to establish that the probability of appearence of a phonetic word beginning from a vowel

(or, on the contrary, consonant) does not depend on what kind of phoneme is in the end of a previous word in a speech chain.

Both above formulated laws allow to derive absolutely theoretically (without any adjusted parameter) elements of the transitional matrix of a markovian chain of the first order, describing alteration of microscopic level marinons and to test numerous sequences, resulting from proposed theory. For instance, it immediately follows from mentioned above regular correspondences that average number of consonants contained in one rhythmon equals 1. Taking into account that the number of syllables in any phonetic word equals total number of rhythmon minus 1, is possible to arrive to the Menzerath-Altmann's law:

$$L(g) = 1 + 1/g \qquad (2)$$

Here L - average length of a syllable (estimated by the number of phonemes), g - length of a phonetic word (in syllables).

Formation of the coherent text is not binded with any single specific mesoscopic level of the text organization. Semantic correlations pierce all the text. Forming of the text as structural integrity is determined by the interdependence of all constituting levels of its totality. The latter allowed to make an assumption that there should be observed some structural likeness between different mesoscopic levels. Taking into account that 1/C equals average length of phonetic words measured by the number of rhythmons contained in them, it is natural to suppose that analogic correspondences should take place also for the number of marinons in any two adjacent mesoscopic levels of the whole text organization: i.e., correspondence between number of syntagms and phonetic words, number of sentences and syntagms, number of paragraphs and sentences etc., should also be characterized by the same universal constant.

Experimental testing of the formulated above hypothesis showed that it is really confirmed with very high degree of precision. Moreover, analysis of marinons' distributions for any two levels i and j permitted to reveal that the distributions posess very high degree of correspondence for any two levels which have the difference between their ordinal numbers i - j = Constant. In this case the length distribution of syntagms, expressed in rhythmons, differs only by chance fluctuations from the distribution of sentences, expressed in phonetic words, and of paragraphs, expressed in syntagms. Just this fact is a principal content of the law of fractal likeness.

Database of Russian Synonyms and its Quantitative-Systemic Analysis

Elena A. Pokrovskaya
Philological Faculty, Moscow State University, Russia

The present paper is concerned with quantitative-systemic investigation of synonyms relations in lexicon in quantitative aspect. A database of synonym groups of minimum and maximum number of units was created on the basis of the "Dictionary of Synonyms" by A.Yevgenieva, which includes 1445 groups of 2 synonyms and 300 groups of 8 and more synonyms. Each unit of a synonym group was scrutinised in relation to the following linguistic parameters: part of speech, stylistic characteristic of the whole word and its lexico-semantic variant, polysemy of the word, existence of homonyms of the word, age of the word, word-formation status of the word, whether or not the word was borrowed, existence of phraseological expressions with the word, frequency. Parameters mentioned above relate to the whole language system, and two more specific parameters - volume of a group of synonyms, if other lexico-semantic variants of the word are units of other synonym groups - were examined.

Each parameter was scrutinised in relation to another parameter, which reveals correlation between them. Some parameters were compared with the data for the whole language obtained from the database which is an excerpt from the "Combined Dictionary of the Modern Russian Lexicon" of R.P. Rogozhnikova.

The source of empirical data for excerpts of synonyms was the "Dictionary of the Modern Russian Literary Language" (first edition - vv. 1 - 17 [1948 - 1965], second edition - vv. 1 - 4 [1991 - 1994], "Word-Formative Dictionary of the Russian Language" by Tikhonov A.N. and a number of etymological dictionaries.

The following results were obtained:

Volume of Groups of Synonyms
As it was said above, marginal groups of synonyms were used for the resent investigation: 2-unit groups and groups of 8 and more synonyms. The maximum number of synonyms forming one group is 33. Table 1 shows the number of groups of different volume:

	Volume of a group\ Number of Groups					
V	2	8	9	10	11	12
	13	14	15	16	18	19
	20	22	23	33		
N	1445	91	61	34	21	13
	11	3	4	1	5	1
	3	1	1	1		

As it can be seen, correlation between the volume and number of groups can be characterised as inversely proportional.

Synonymic activity:
Though more than 50% of synonyms in each of our excerpts of synonyms are used in one group of synonyms only, more synonymically active words, i.e. words which lexico-semantic variants are used in several different groups of synonyms, are concentrated in Excerpt of maximum volume groups.

Parts of Speech:
According to the dictionaries, the whole language is characterised by the predominance of nouns, followed by verbs and adjectives, which are the three main parts of speech in the Russian language (totally around 94% of all the words registered in the dictionaries). Percentage for all parts of speech for 2-unit groups is almost the same as those in the Dictionary of the Russian Language, while percentage for the excerpt of groups of maximum volume shows predominance of verbs and adjectives over nouns. Therefore, non-attributive part of speech (nouns) tends to form 2-unit groups rather than groups of maximum volume; while attributive parts of speech (verbs and adjectives) are more characteristic of groups of maximum volume.

Polysemy
Words forming synonym groups of any volume are more polysemantic than words of the whole language. Thus, average polysemy in the whole language is 1.67 meanings per word, while average polysemy of the 2-unit groups of synonyms is 2.95, and of the groups of maximum volume - 2.63 meanings per word.

Age of Word
In order to make comparison easier, ages of words were grouped into 7 periods. In total, the percentage of indo-european, old Slavic and ancient Russian words forming groups of synonyms is higher than the relative percentage of the words in the whole language. The highest percentage of words in both excerpts of synonyms relates to 18th century (35 - 36%), while in the excerpt of words of the whole language the peak related to words of the 19th century (around 30%).

Frequency:
According to calculations made on the basis of excerpt from the whole language, the average frequency of the words of the whole language is 14.59; average frequency of excerpts of synonyms is much higher - 52.2 for 2-unit groups of synonyms and 41.3 for groups of synonyms of maximum volume. As the difference here is substantial, conclusion can be drawn that relatively high frequency is characteristic of words forming groups of synonyms.

Stylistic Characteristics:
This parameter opposes two excerpts of synonyms: the overwhelming majority of lexico-semantic variants or words of 2-unit groups are stylistically neutral, while more than 50% of synonyms in groups of maximum volume are stylistically marked with strong predominance of colloquial words.

Derivational characteristics:
Derivative words prevail in both excerpts, root words tend to form 2-unit groups rather than groups of maximum volume, and composite or composite-derivative words are more characteristic of 2-unit groups.

Borrowings:
Borrowed words tend to form 2-unit groups, which, in our opinion, correlates with predominance of special words in the excerpt and their tendency to form groups of 2 synonyms.

Homonyms:
The share of words having homonyms is low and is almost the same in both excerpts.

Phraseological expressions:
Almost 1/5 of words in each of the excerpt have phraseological expressions.

Text as real population in A.A.Chuprov sense

Sergei V. Chebanov
31 Moika, Apt.12, St.Petersburg 191186, Russiaa
E-mail: cheb@cc.ief.spb.su
S.M. Sechenov Inst. of Evolutionary Physiology and Biochemistry
(Russian Academy of Sciences, St.Petersburg)
Institute of national model of economy (Moscow)
Gregory Ja. Martynenko
35 12th Line, Apt.59, St.Petersburg 199178, Russia
University of St.Petersburg, Dep. math. and struct. linguistics.

Interpretation of statistics application results in linguistics and philology requires clear understanding of logic status of used instruments, primarily that of analysed material itself.

It is now generally accepted to think concrete text to be collective notion representing internal system. On the contrary, lexicon, in particular that of a given text, is thought to be distributive notion represented by external system. As noted by Yu.A.Shreider, external system here appears as a class of states or an aggregate of components of internal one.

Said interpretation is now a common place, the more so as being in accordance with discrimination between collective and distributive sets, which is of great importance for surmounting of B.Russell's paradoxes in set theory. The alternative approach is developping S.Lesniewski mereology as collective sets theory as opposed to "naive" set theory as distributive sets theory. In developping the above ideas S.M.Meyen proposed the opposition of meronomy and taxonomy.

The resulting tree of notions may be represented as follows:

Categories	Collective	Distributive
Notion	collective	distributive
Volume of notion	semantic size (class of details)	class of referents
Relationship of volume and content	direct	inverse
Procedure of partitioning	articulation	division
Incomplete basis induction problem	reconstruction	extrapolation
Components	heterogenous	homogenous
Representing system	internal	external
System	organized in time and space	abstract logical construction
Component aggregate (population)	collective set	distributive set
Model	Lesniewski mereology	Cantor set
Probabilistic concept	mathematical statistics	probability theory
Examples	text, 19th century Russian literature	lexicon, list of works

Within this context statistical notions as such were never specifically considered, though the opposition of mathematical statistics and probability theory discussed.

Therefore the works of A.A.Chuprov are of interest, who focused on logical basis of statistics early in this century.

As an exact equivalent of the opposition of collective and distributive notions, he had that of group and generic notions.

On this basis, he introduced discrimination between two types of statistic aggregates, one of which he named real population and the other descripted as artificial without giving them any special name.

Detailed examination of the notion of real population leads to conclusion that this is not simply statistical collective notion, but the type of collective notion, to which rigid, not soft systems correspond, internal, as determined much later by A.A.Malinowski (son of A.A.Bogdanov, originator of tectology).

A.A.Chuprov's approach revealed characteristic features of collective notions of statistical nature.

In this context, it is very important that A.A.Chuprov entered the discrimination under consideration into problematics of the Rickert's opposition of ideographic and nomotetic knowledge by demonstrating that statistical knowledge is intermediary type of knowledge between two Rickert's types of knowledge.

Panchronic and panchorologic features of distributive notion reveal its nomotetic nature, while temporal and locative definiteness of collective notion indicates its ideographicity (here we have, however, some problems with collective notions that have not a single empirical referent, but are rather generalized structure of a class of referents).

The idea of A.A.Chuprov is extremely interesting that on the basis of such an interpretation of ideographic and nomotetic knowledge it becomes surmountable the diadicity of opposition of fundamental types of notions, which gives the way to introduction of two more classes of notions (A.A.Chuprov had not quite clear terms):

ideographic generalizations, for which notions locative parameters are fixed and temporal parameters are not, and

relative historical notions, fixed in time and not in space.

Then along with collective and distributive notions in linguistic studies four possible classes of aggregates should be discerned, as may be illustrated in the following examples:

Collective notion: story as a literature form opposed to novel, tale, etc, or "Chameleon" by A.P.Chekhov.

Distributive notion: second half of the 19th century Russian stories.

Ideographic generalization: Russian story in general, from N.M.Karamzin to future end of the form.

Relative historical notion: 1890's stories written at any place, possibly including extraterrestrial locations.

Discerning the nature of a notion is important in carrying out concrete linguostatistical studies, as each type of aggregates requires specific procedures of sampling, representativeness determination, and results interpolation.

Summary
Topical paper
Text presents through collective categories (notion, set), and dictionary - through distributive one. In A.A.Chuprov's sense stochastic sets represented by real population (collective notions), artificial aggregates (distributive notions), ideographic generalization (sets not certain in time) and relative historical notion (not certain in the space). All four types of sets exist in linguostatistical studies.

Model Thinking as the Qualitative Foundation of Linguistics

Jouko Seppänen
Helsinki University of Technology
Department of Computer Science
FIN-02150 Espoo, Finland

In system theory one distinguishes between two classes of systems, namelyobject systems and model systems, i.e. systems standing for or representingsomething else than what they themselves actually are. Subcategories ofmodel systems are signs, codes and languages, including natural languagesas well as thought. In this paper we will review the history and philosophyof model thinking and its central notions, including concepts like object,subject, feature, relation, analogy, homology, classification, metaphoretc. and consider model theory as a conceptual and qualitative foundationfor theoretical and quantitative linguistics.

The use models of various kinds is common in everyday life and in science.The principles associated with and problems arising from using models havebeen studied in many fields: in philosophy and methodology of science,physics and mathematics and other sciences as well as in art andengineering. Special modelling and simulation techniques have beendeveloped in analog and digital computing, information and computer scienceand other fields. Related questions are discussed in theoreticallinguistics and semiotics but still there is no well established andgenerally recognized theory of models or model philosophy. Such a scienceshould make explicit the theoretical foundations of using, developing andinterpreting models and languages and the validity of knowledge thereby obtainable and represented.

The theory of models has a close connection with functionalism.Functionalism is a philosophical view which maintains that functions andqualities can be relatively independent of specific underlying media ormechanisms in which they are realized and that the same functionalities canbe realized in different ways. In philosophy this question is exemplifiedby the mind/body problem.

Understanding and defining the principles and notions involved in analogyand metaphor precisely it becomes possible to define precisely also thenotion of information as a fundamentally subjective concept. A subjectivetheory of information is necessary to define precisely higher level notionslike knowledge, thought, language, communication, control, measurement,mental image, aim, goal, interpretation, meaning, context, world model etc.which are central to analysis of mental, linguistic and cultural systemsand processes including art, religion, philosophy and science.

Models can be characterized in terms of degree of correspondence andconfidence with respect to the object being modelled. The degree ofsimilarity can vary from partial resemblance to an identical copy - illustration, lat. in + lustre, lumen, light, throw light,elucidate, exemplify

- visualization, lat. visu, sight, make visible, form a mental picture
- simulation, lat. simil, same kind, assume same appearance

– emulation, lat. emul, rival, strive to equal
– imitation, lat. imago, image, behave or become alike
– animation, lat. anima, soul, give life to
– realization, lat. res, thing, make according to a plan
– copying, lat. copios, abundant, make a similar exemplar
– self-reproduction, make a copy of oneself by oneself.

Analogies allow inferences to be made about the target system on the basisof what is known of the model. An analogy may be partial, when it is asimilarity relation, or complete, when it is an identity or one-to-onerelation. Usually models are partial, since a complete model is a copy. Ananalogy can be physical or formal. In physical analogy there is similarityof property, form, structure or function whereas formal analogies may belogical, mathematical or computational. A model may be actual, mental orsymbolic, i.e. an object, a thought or a verbal or sign model, and it canstand for other actual, mental or symbolic objects as targets. Sign modelsare traditionally subdivided into pictorial, diagrammatic and textual. Insemiotics sign models are subclassified as iconic, indicial and symbolic.

References

1. Seppänen J. (1997): History and Philosophy of Human and Social Sciences. In: Altmann G., Koch W.A., eds., Systems: New Paradigms for the Human Sciences, de Gruyter, Berlin.

PSYCHOLINGUISTICS

LANGUAGE

ACQUISITION

THE CHI-SQUARE TEST AND ITS SIGNIFICANCE IN STUDYING STABILITY IN RESPONSE PATTERNS

Amitav Choudhry
Linguistic Research Unit
Indian Statistical Institute
203 B.T. Road, Calcutta - 700 035, India
e-mail : chou@isical.ernet.in

SUMMARY

The paper examines the significance of the Chi-square test in studying stability in response patterns in language attitude questionairres which are based on the Likert method. The paper also tries to justify the validity of adopting the 3-point opinion scale instead of the 5-point scale. The paper is based on an empirical study on the language attitudes of Bengali (Indo-Aryan) speaking subjects in the context of Indian plurilingualism.

ABSTRACT

In sociolinguistic research there are various means of collecting information which may subsequently be subjected to either qualitative analysis or quantitative analysis or both. For any kind of quantitative investigation we need ways of making sense of the data and this is the purpose of statistical methods. According to Butler (1985) courses in the application of statistics concentrate far too heavily on the methods themselves and sufficient attention is not paid to the reasoning behind the choice of particular methods. Most linguists are not interested in the more theoretical side of the application of methods nor the mathematical mode necessary for the derivation of formulae. Where the urgency to apply statistical methods is felt, one just wants a menu-based application which can take care of quantification of a given set of data. They are more anxious to fit their data into a given format and in the process monitor their research over a tailor-made path so that some quantitative deductions can be made, with scant respect to whether the final figures which may even look fanciful, do justice to the analytical aspects of the research in question. It is important for researchers to understand the rationale behind a particular method before attempting to apply it on a given set of data.

Once we have a set of data, sometimes it is necessary to draw conclusions after carefully examining every occurrence, or every response from a chosen

phenomena. Similarly we are often concerned not with the characteristics of just one set of data, but with the comparison of two (or more) sets. For example we may be interested in testing the hypothesis that a group of informants who have been chosen for a particular survey, meet the stability criteria in their response patterns, and this conclusion can only be drawn after a comparison is made between their responses to a given set of statements which are part of an attitude questionairre. This in turn should be statistically verifiable to justify the significance of uniformity in their response patterns.

When comparisons are involved, we need to know not only the general characteristics needed to coin a statement but also whether the characteristics of the two statements are sufficiently similar or different for us to come to any plausible conclusion, which can be stated factually as representive views of the respondents in question. In the section on empirical evidence we will see the importance of coining appropriate statements which in turn is more likely to reflect the attitudes of respondents in a given context.

A study of the language attitudes of Bengali (Indo-Aryan) speaking subjects (N=200) in Hyderabad (South India) was conducted by me. Though the permanently settled Bengalis had a more accomodative attitude towards Telugu (Dravidian and also the dominant regional language), a majority of them did not consider learning the regional language more important than cultivating the mother tongue. English (Associative official language) was preferred to Hindi (Primary official language) as the medium of instruction in schools and for higher education. It was also found that their attitude towards English was highly positive on many counts. What I have done here is taken a few statements which were part of the attitude questionairre used in my study and for the purpose of computing Chi-square, compared a few pairs of statements to see the stability in the response patterns of the subjects in question. The study concludes that the claimed attitudes of the respondents were fairly uniform with a marked stability in their response pattern and also emphasizes the need for such studies in language planning strategies in multilingual India.

Topical paper; Topic Area : Applications of methods

Native Speakers' Reactions to Modern English Usage

Peter Kunsmann
Freie Universität Berlin
Institut für Englische Philologie
Goßlerstraße 2-4
D-14195 Berlin
wpkuns@zedat.fu-berlin.de

Johannes Gordesch
Freie Universität Berlin
Institut für Soziologie
Babelsberger Str. 14-16
D-10715 Berlin
jgord@zedat.fu-berlin.de

Burkhard Dretzke
Freie Universität Berlin
Institut für Englische Philologie
Goßlerstraße 2-4
D-14195 Berlin
dretzke@zedat.fu-berlin.de

Summary: The paper deals with the question of usage in modern American English. A questionnaire with sentences of disputable correctness was sent to randomly chosen people in the US and was then analyzed statistically and interpreted linguistically. A graph theoretical clustering algorithm is developed in some detail and forms the basis of the analysis.
Topical paper: *Sociolinguistics*

1 The Empirical Study

1.1 Selection of Items for the Study

The discussion of correct or appropriate usage has a long tradition. Over the years a list of items has been accumulated that serves as one source for our investigation. Other items have been added to this list from personal observations of the authors. Eventually, authentic data from word corpora will also be included. For the present study, selected items were combined in a questionnaire for presentation. Subjects were required to perform three separate tasks. Thus, section I contained randomly placed statements of *divided usage*. The subjects had to react to these statements labeling them 'correct' or 'incorrect'. In section II the subjects were asked to supply the appropriate *question tag* to a given statement. In the final section, the subjects were presented with two, three or four statements per item and were asked to rate the different statements as to their relative correctness. The version of the questionnaire reported on here contained a total number of 71 items. Only items from Section I (48 items) form the basis for this report.

1.2 Selection of Subjects for the Study

In the literature, a number of studies have relied on anecdotal or accidental evidence. Most subjects were members of a small academically oriented group of *educated native speakers*. In order to start from a broader base of socially relevant groups, this study, in contrast, relied on a random mail poll in a number of different cities in the United States: Washington, D.C. and Philadelphia in the East, Ann Arbor and Milwaukee in the Midwest and Pittsburgh and Cleveland half-way between these points. In order to analyze social variables, the subjects were asked to mark their age, gender and education, and optionally their occupation. Approximately 1200 questionnaires

were distributed to 600 postal addresses. 207 of these were returned. For the purpose of evaluating the questionnaire, the subjects were divided into three age groups (20-30, 30-50, above 50) and into four groups according to education (high school diploma, some college, college graduate, advanced degree).

1.3 Results

With regard to both the social variables and the items of divided usage some interesting results can be reported.

Looking at the correlation between age and gender, it can be observed that the coefficient for subjects above 50 years of age is .96, followed by the 30-50 year-olds with a coefficient of .91. The lowest correlations are observed in the groups 'women above 50' vs 'men 20-30' with a coefficient of .72 and 'men above 50' vs 'women 20-30' with a coefficient of .73. Comparing men and women on the variable of educational background, the highest correlation is observed for college graduates (.94), followed by those who earned an advanced degree (.92) and 'male college graduates' with 'female advanced degrees' (.91). The lowest correlation is observed with high school graduates (.52) and 'female with some college' vs 'male high school graduates'.

An interesting item on the questionnaire is the one dealing with the use of subjective pronouns in coordinate constructions after a preposition. Eight of the items on the questionnaire dealt with this divided use of *I* and *me*.

(1) There is only one man between he and the goal line.
(2) Between you and I, our neighbors drink heavily.
(3) He came after Alan and I, and he shot him.
(4) John invited Bill and I for dinner.
(5) She told Charles and I the whole story.
(6) I think it is up to you and I to decide.
(7) Is that the kind of world God intended you and I to live in?
(8) It's about time for John and I to buy a new house.

These statements may be grouped into four grammatical structures. (1) shows the pronoun preceding the lexical noun phrase in the coordinated prepositional phrase; (2) and (3) are prepositional phrases where the coordinated objects are both pronouns; (4) - (6) require objective case on the basis of the coordinated direct lexical and pronominal objects; and (7) - (8) constitute utterances in which the coordinated noun phrase is the subject of an infinitival complement sentence.

Preliminary results on the basis of 207 questionnaires show the following distribution of acceptability:

(1) -- (18 acceptable - 182 not acceptable)
(2) -- (54 - 151)
(3) -- (17 - 186)
(4) -- (53 - 147)
(5) -- (50 - 152)
(6) -- (56 - 145)
(7) -- (52 - 149)

(8) -- (70 - 133)

1.4 Discussion

There appears to be a clear distinction between the low acceptability of (1) and (3), the high acceptability of (8) and the rest of the sentences. The low acceptance of (1) and (3) is explained by the fact that they contain a lexical noun in the coordinated structure. For the other sentences a number of linguistic arguments can be advanced. The structural complexity of the utterance, the phenomena of lexicalization and change of grammatical categories as well as pragmatic considerations determine the relative acceptability of these items. In (2), for instance, the coordinated pronouns are felt to be a single linguistic unit with the status of a *standing phrase*.

The results for items (2) and (4) to (7) show only insignificant differences. At first glance, therefore, the linguistic division made above seems to have no affect on their acceptability. However, upon closer examination, the weight of the linguistic arguments may be different for the individual items.

Sentence (8) can be accounted for by an analysis of complement sentences in English. Finite complement sentences are formed with the conjunction *that*. Therefore, the finite sentence corresponding to (8) may be (9):

(9) It's about time that John and I buy a new house.

Thus for native speakers accepting (8) the preposition *for* is considered as a conjunction when the embedded sentence is non-finite in analogy to the conjunction *that* in finite sentences.

The linguistic arguments presented here have to be supplemented by non-linguistic ones such as affective variables, social and situational factors, and political or cultural backgroungs of individual speakers.

A number of conclusions can be drawn on the basis of this study. First of all, divided use of linguistic items may be accounted for by linguistic considerations. Secondly, the complexity of the utterance determines in part its acceptability. And finally, non-linguistic factors will have to be taken into consideration in an analysis of items of *divided usage*.

2 Mathematical Models

Determining a partition of an assemblage of objects into maximal collections of suitably similar objects is, apart from its analysis, one of the main tasks in statistics. From the linguist's point of view, however, the elements to be combined into linguistically relevant subclasses are structural objects rather than mere n-tuples of numbers. Algorithms, therefore, should reflect this and operate on finite structured sets.

2.1 Categories and Functors

In an abstract way, survey data are sets with certain structures studied along with a class of mappings that preserve these structures. Thus the concepts of category and of functor provide a working tool for mathematical and statistical analysis. A *category* K consists of two classes (not

necessarily sets) O_K and M_K, for which the members of O_K are called *objects*, the members of M_K are called *morphisms*, and the following conditions are satisfied:

C1 With each ordered pair (a, b) of objects there is associated a set M_K of morphisms such that each member of M_K belongs to exactly one of these sets;

C2 if f is in $M_K(a, b)$ and g is in $M_K(b, c)$, then the composite $g \circ f$ of f and g is defined uniquely and is a member of $M_K(a, c)$;

C3 if f, g, and h are members of $M_K(a, b)$, $M_K(b, c)$, and $M_K(c, d)$ respectively, so that $(h \circ g) \circ f$ and $h \circ (g \circ f)$ are defined, then $(h \circ g) \circ f = h \circ (g \circ f)$;

C4 for each object a, there is a morphism e_a in $M_K(a, a)$, called the identity morphism, such that $f \circ e_a = f$ and $e_a \circ g = g$ if there are objects b and c for which f is in $M_K(b, a)$ and g is in $M_K(a, c)$.

Further useful concepts are zero morphism, isomorphism or equivalence, automorphism, etc. In a similar manner, the concept of a functor mapping category K into category L is introduced. In particular, covariant functors, contravariant functors, isomorphisms and anti-isomorphisms are defined.

Let K and L be two categories with O_K , M_K and O_L , M_L denoting, respectively, the classes of objects and morphisms in K, and the classes of objects and morphisms in L. A *covariant functor* of K into L is a function F whose domain lies in the class of all objects and morphisms in K; which maps O_K into O_L and, for each a and b in O_K, maps $M_K(a, b)$ into $M_L[F(a), F(b)]$; and which has the properties:

F1 If e_a is the identity morphism in $M_K(a, a)$, then $F(e_a)$ is the identity morphism in $M_L[F(a), F(b)]$;

F2 if f and g are morphisms in $M_K(a, b)$ and $M_K(b, c)$ respectively, then $F(g \circ f) = F(g) \circ F(f)$.

A *contravariant functor* is defined correspondingly, except that F maps $M_K(a, b)$ into $M_L[F(b), F(a)]$ and the equality in F2 is replaced by $F(g \circ f) = F(f) \circ F(g)$.

2.2 Clustering Structures: Graph-Theoretic Clustering Algorithms

Starting with these abstract concepts, clustering algorithms are developed to obtain a classification into meaningful classes.

A cluster is a maximal collection of suitably similar objects drawn from a larger collection of objects. A combinatorial cluster analysis model is appropriate where either the raw data are in the form of a similarity relation or where the number of objects is too large for distance matrix methods to be computationally tractable. Objects are presented by vertices of a graph, and those pairs of objects satisfying a particular similarity relation are termed adjacent and constitute the edges of the graph. Clusters are then characterized by appropriately formed subgraphs.

To put it more precisely: Given a set $S = \{o_1, o_2, \ldots o_p\}$ of objects, one can define a non-negative real-valued 'proximity' function F on S×S where $F(o_i, o_j)$ for $1 \leq i, j \leq p$, is a number (in the more general case an element of an ordered Archimedean group) measuring the 'similarity' of the two objects o_i and o_j. $F(o_i, o_j) < F(o_i, o_k)$ indicates that the objects o_i and o_j are more similar than the objects o_i and o_k. Given such a function F and a number s, a graph G(s) can be defined where the set of vertices $V(G(s)) = S$ and $edge(o_i, o_j) \in E(G(s))$, the set of edges of the graph G(s), if and

only if $F(o_i, o_j) \leq s$.

The probability that a given set T has the similarity structure introduced in the graph G(s) can then be investigated by comparing the graph G(s) with all other graphs having the same number of vertices and edges as G(s).

2.3 Computational Aspects

Computer Algebra Systems (CAS) are systems 'for doing mathematics'. In particular, they provide useful tools for constructing and manipulating graphs. *Mathematica* supplies the package 'DiscreteMath', and *Maple* the packages 'combinat', 'comstruct', and 'networks'. While CAS allow for linguistic as well as mathematical modeling, specialized packages for cluster analysis are preferable from the computational point of view.

References

Linguistics

The American Heritage Dictionary (1969) New York
CRYSTAL, D. (1985) To Use or not to Use. *English Today* 1
DRETZKE, B. (1992) Neuerungen in der englischen Sprache - Divided Usages. *Fremdsprachenunterr.* 2
GORDESCH, J. and B. DRETZKE (1997) Correctness in language - a formal theory. In: G. ALTMANN et al (eds.), *Linguistic Structures. To Honor J. Tuldava* (to appear)
HONEY, J. (1995) A new rule for the Queen and I. *English Today* 2
KUNSMANN, P. (1995) Grammatikalität und Akzeptabilität im amerikanischen Englisch. *Fremdsprachenunterricht* 6
MATHEWS, M. (1963) *The Beginnings of American English.* Chicago
QUIRK, R. and J. SVARTVIK (1966) *Investigating Linguistic Acceptability.* The Hague
Random House Webster's College Dictionary (1991). New York
SHERWOOD, J. (1960) Dr. Kinsey and Mr. Fries. *College English* 21
Webster's Third International Dictionary (1961). Chicago

Mathematics and Statistics

KILLOUGH, G. and R. LING (1976) *JASA* 71, 213-300
MATULA, D. W. (1977) In: J. van RYZIN (ed.), *Classification and Clustering.* New York, 95-129
REDFERN, D. (1996) *The Maple Handbook. Maple V Release 4.* New York
SKIENA, St. (1990) *Implementing Discrete Mathematics. Combinatorics and Graph Theory with Mathematica.* Redwood City, CA
WOLFRAM RESEARCH (1996) *Standard Add-on Packages.* Champaign, Ill, and Cambridge, UK

ASSOCIATIVE LINGUISTIC EXPERIMENT AND ELABORATION OF METHODS OF COMPUTER DIAGNOSTICS FOR HUMAN'S INBORN-HEREDITARY SYNDROMES

Vladimir A. Dolinsky, Ph.D.; Sergey S. Rudakov, M.D., Sci.D.

V. Dolinsky, Moscow State Linguistic University. 129345, Moscow, Ostashkovskaya, 9-2-98. Russia. Tel. (095) 475-8384. E-mail: nalimov @ Nalimov.home.bio.msu.ru

S. Rudakov, The Pirogov Russian Medical University. 123458, Moscow, Tallinnskaya, 9-2-124. Russia. Tel. (095) 942-0815.

The original project of developing diagnostic programm for syndromology based on results of psycholinguistic word association test, its quantitative analysis and construction of data-base modelled on probability semantic networks.

PROJECT NOTE.

Automated systems, psycholinguistics, methodology, clinical genetics, syndromology, quantitative methods.

1. The universal nature of modern methods of diagnostics and treatment for human diseases and development defects is well known. The same goes for clinical polymorphism of these diseases and development defects, and this demands individualization of diagnostic and treatment approach. Up to now it has been carried mainly out on an empirical basis. This si-

tuation has serious drawbacks which may be eliminated with the help of the synthesized syndromological and psycholinguistic approach (as they use quantitative methods) that we suggest for the solution of applied problems of clinical medicine.

The essence of syndromological approach consists of revealing the syndromal spectrum (rank-frequency distribution of inborn-hereditary syndromes) of human diseases and development defects with the subsequent group analysis of clinical polymorphism and the development of a group treatment and diagnostic approach. The theoretical background of the syndromological approach is based on the assumption that the inborn-hereditary syndrome is a clinically significant variant of human constitution. The hereditary nature of most syndromes, on the one hand, and the fact that they reflect development faults, on the other hand, allows us to consider these syndromes as snapshots of clinical polymorphism. At the same time, the assumption of polygene and expressive gene action, based on the theory of "channel development", makes it possible to use the data obtained as the result of observing syndromal forms of diseases and human development defects, in analysing "incomplete" syndromes and isolated development defects and diseases.

2. We think that all these difficulties may only be overcome by using computer methods of diagnostics. We made the first step in this direction by developing the "Diagnostic Point" (DP) program. Its diagnostic algorithm is based on the available statistical data and on the idea of diagnostic significance of the symptoms of inborn-hereditary syndromes. This significance reflects the expectations of the specialist working with the system.

The London Dysmorphology Database (LDDB) and the London Neurogenetics Database (LNDB), which were both developed by Dr M.Baraister and Dr R.Winter (Institute of Medical Genetics, University of London, UK) and the P.O.S.S.U.M. which was prepared by Dr. Agnes Bancer (The Murdoch Institute for Research into Birth Defects. Malburn, Australia) as tools for the diagnosis of multiple congenital defects and inborn-congenital syndromes. The DP system as a tool for the diagnosis of inborn-hereditary syndromes as well.

The LDDB and LNDB are typical data base systems, without diagnostic functions while the P.O.S.S.U.M and DP-system have a diagnostic function. The DP diagnostics algorithms are based on the available statistical data and on the idea of diagnostic significance of the symptoms of inborn-hereditary syndromes. This significance reflects the expectations of the specialist working with the system.

Notwithstanding the obvious practical benefit of this programmes, we recognize its limitations. We therefore suggest the following project, which involves development of diagnostic programme for syndromology of a human being, based on the results of well-khown psycholinguistical experi-

ment - free word association tests.

3. The need for this approach follows from the fact that the above mentioned reasons for the drawbacks of syndromal diagnostics may be formulated as semantically indistinct description of syndromes. This is due to insufficient knowledge of semantic relations between these descriptions and features of the syndromes and also to the lack of understanding of their linguistic structure.

Associative potential of a terms is the least studied of vocabulary semantic characteristics. Its investigations rests upon the test well known in the psycholinguistics. It consists in showing a word stimulus to a person and asking him to respond to that stimulus with the first word that came to his mind. In the word association test in which single answers (discrete associations) of a group of respondents (subjects) are registred, the response parameters are diversity (availability) and frequency of occurence.

The frequency of a particular response reflects the currency of a given association in the given group and is indicative of identical connections between the stimulus and response occuring of the minds of different people. The availability of different associations received from a group of respondents (subjects) indicates the broadness or narrowness of the associative spectrum engendered by a given word stimulus.

Reactions produced by respondents can be regarded as

more or less typical of the given stimuli, specific for the given speech community (experts in different fields of clinical medicine), and regularly recurring in repeated tests.

4. We expect that indistinct description in syndromology may be essentially limited if not eliminated. As a result, the list of inborn-hereditary syndromes and their indicators may be verified by applying the controlled associative experiment. In this experiment, experts in syndromology and other fields of clinical medicine will have to give verbal responses to word-stimuli related to the informative signs of various syndromes. The rank-frequency and the spectrum-frequency distribution of these associations will reflect qualitative and quantitative parameters of "semantic fields" of syndromes.

We plan to conduct a controlled associative experiment in order to register discrete and/or continuous response to the offered stimuli and present the obtained data with the help of special software imitation of collective associative memory. It is also expected that the associative experiment will give us the solution to one of the key problem of artificial intelligence - integration of the information obtained from different experts.

The methodology and the mathematical apparatus of the associative data processing have been elaborated by V.A.Dolinsky in detail]. The wave structure of verbal associations identified recently is of particular interest for the deve-

lopment of algorithms. The data of the linguistic associative experiment will be used for the construction of the "Associative Thesaurus of Syndromes" data-base modelled on probability semantic networks. The system "Associative Thesaurus of Syndromes" will make it possible to carry out differentiated diagnoses of human inborn-hereditary syndromes.

This data based on the above mentioned algorithm will provide the user with information on both syndromes and symptoms.

5. Research plan envolves:

- elaboration of a list of "word-stimuli" (symptoms of inborn-hereditary syndromes).

- elaboration of a list of experts (about 100-150 experts in the syndromology and other fields of clinical medicine).

- realisation of associative experiment (distribution of word responses and collection of data).

- the computer processing of the experiment's data (Associative Thesaurus of syndromes).

- elaboration and realisation of associative algorithms of diagnostics of inborn-hereditary syndromes.

- unification of P.O.S.S.U.M., LDDB, LNDB and DP on the base of associative diagnostics algorithms, elaboration of a new diagnostics system for syndromology.

- clinical examinations of this system.

Is forbidding not allowing? And why not? A meta-analysis

Bregje C. Holleman
Uil OTS - Utrecht University
Trans 10, 3512 JK Utrecht, The Netherlands
e-mail: b.holleman@let.ruu.nl
Qualico'97 - Preliminary Paper (topical)

0. Summary

topic areas: wording effects, congenericity, meta-analysis, question/answer process

Wording effect research gives the opportunity to investigate question/answer processes and differences in meaning in a relatively natural task. The forbid/allow asymmetry is a wording effect that has received a lot attention. Although the verbs 'forbid' and 'allow' are generally considered each others' counterparts, the answers to attitude questions using those verbs are not each others' opposites. In this paper it is demonstrated how a meta-analysis gives more insight into explanations for this wording effect in a way that facilitates generalisability of the findings. Research reported here leads to three conclusions: 1) the forbid/allow asymmetry exists all in all; 2) the attitudes measured with forbid/allow questions are the same, but they are expressed differently on the answering scales due to the use of those verbs; 3) part of the explanation for the asymmetry might be the strong connotations of both verbs, but the answering behaviour of indifferent respondents, or respondents with a weak or very subtle attitude, also seems to be an important factor.

I.1 The forbid/allow asymmetry

Reseach has repeatedly demonstrated that small changes in the wording of a question cause huge differences in the responses obtained (see Jobe & Mingay, for a review). This raises questions about the validity of survey questions: which particular question wording measures what the questionnaire designer intends to measure? The basic goal of research into wording effects is traditionally to generate practical advice for questionnaire design. At least equally important, however, is the more fundamental goal to gain insight into the cognitive processes underlying question answering, and into the variables that affect responses (Cicourel, 1982). Wording effect research gives the opportunity to investigate question/answer processes and (differences in) meaning in a relatively natural task, thus providing ecological validity.

A wording effect that has received a lot of attention for more than half a century of research is the forbid/allow asymmetry, first identified by Rugg (1941). Although 'forbid' and 'allow' are considered each others' counterparts, the answers to questions with the verb 'forbid' turn out not to be opposite to the answers to questions with the verb 'allow'. Rugg found that respondents were more likely to support freedom of speech when the question was worded with the verb 'forbid', than when it was worded using the verb 'allow' - resulting in a difference of 21% between answers to two questions that are generally considered to be logically equivalent (see Table I).

The forbid/allow asymmetry is not always found. When it is not, researchers tend to formulate post-hoc hypotheses to explain the absence of the effect. Those hypotheses are hardly ever tested. Neither is it obvious whether this would be worthwile: most forbid/allow experiments consist of one manipulated question only, so there are always several possible causes for the asymmetry not to occur: differences in respondent characteristics, as well as administration mode, question issue, et cetera (Waterplas et al., 1988).

Do you think the US should forbid public speeches against democracy?
yes, forbid 54%
no, not forbid 46%
Do you think the US should allow public speeches against democracy?
yes, allow 25%
no, not allow 75%

Table I: Forbid/allow Experiment by Rugg (1941)

The fact that most experiments only consist of one manipulated question, and that analysis is

always done over only one question, also causes difficulties generalising the wording effect. On first sight, looking at all the experimental results, it is not certain whether the asymmetry exists all in all.

Therefore, a meta-analysis of all forbid/allow research was conducted: first of all the goal was to find out whether the asymmetry exists overall, secondly the goal was to explore whether the posthoc hypotheses are supported when analysing over all the forbid/allow questions reported in literature. Methods and results of the exploratory part of the meta-analysis will be discussed in section II.

The first part of the meta-analysis was done using a (sort of) t-test over all forbid/allow questions found in literature (52) at the same time. The answers were weighed based on the number of respondents per question, so that questions answered by 1500 respondents weigh heavier than questions answered by 40 respondents.[1] Results of the first analysis show that the wording effect does exist all in all. The mean size of the asymmetry was 14%: the answer 'no, not forbid' is given 14% more than the answer 'yes, allow' ($p<.001$). The variance was huge however (sd 9.85), indicating that the size of the asymmetry differs substantially over questions and over experiments.

I.2 Explanations for the forbid/allow asymmetry

The basic explanation for the forbid/allow asymmetry, point of departure in all forbid/allow literature, focuses on the *connotations* of 'forbid' and 'allow': "the former sounds harsher and may therefore be more difficult to endorse, whereas the latter in some context might seem to encourage a deviant behavior and therefore may invite opposition" (Schuman & Presser, 1981:296). "Thus what we have called tone of wording, could be the sole source of the effect." (Schuman & Presser, 1981:280).

Although it does seem plausible for the connotations of 'forbid' and 'allow' to cause the asymmetry, some problems are attached to this explanation. First of all, the explanation as worded by Schuman & Presser predicts for the asymmetry to occur quite constantly, and always have about the same size. The amount of variance found in the first part of the meta-analysis (see section I.1) raises serious doubts on this point. In section II, the second part of the meta-analysis of forbid/allow research indicates that the connotations of 'forbid' and 'allow' are not the only cause of the asymmetry, as the size of the asymmetry turns out to depend of several question content and respondent characteristics.

Secondly, the explanation does not provide any real insight into the cognitive mechanisms underlying the asymmetry, causing it to remain a hypothesis rather than an explanation. For it is not clear in which stage of the question/answer process the asymmetry is theorised to occur - in the stage of attitude localisation, or in the stage in which the respondent maps his/her answer to one of the precoded answer categories. Does the explanation, as worded by Schuman & Presser, mean that answers to 'forbid' questions reflect different attitudes than answers to equivalent 'allow'- questions? The connotations, or semantic fields of both verbs in general, might be that strong that not only the attitude towards a specific issue (abortion, for example) is measured, but also a general attitude towards forbidding or allowing. But it may also be the case that the asymmetry results from slight changes in perceptions of the meanings of attitude questions response options. Krosnick & Schuman (1988) theorise the asymmetry to be caused by differences in the way respondents map their answers to the answering options due to the use of 'forbid' and 'allow': "[...] 'not allowing' is perceived as a less extreme stance than is 'forbidding'."

Whether the forbid/allow asymmetry is caused by the retrieval of (partly) different attitudes caused by the use of both verbs, or whether it is caused by a difference in mapping of the answers to the answering options, was tested by setting up two experiments (one on attitudes towards environmental issues one on attitudes towards ethnic groups) using a correlational

[1] Analysis was done using a method of IGLS, distinguishing between two levels of variance: variance between questions within experiments, and variance between experiments. This was done because questions within the same experiments are more similar to each other than questions administred in different experiments (different respondents for example). Distinguishing two levels made it easier to interpret the results. However, here results will be reported using the overall variance figures.

design. Data were analysed using a structural modelling approach.

Analysis revealed 'forbid' and 'allow' questions were congeneric (Jöreskog, 1971). This means that ranking of respondents based on their answer to 'forbid' questions results in a similar order of respondents as would a ranking on the basis of their answers to 'allow' questions. This may lead to the conclusion that questions worded with either verb do measure similar traits, similar attitudes. However, in both experiments observed scores and error scores to 'forbid' questions differ from those of 'allow' questions: similar attitudes are expressed differently on the answering scale due to the use of 'forbid' and 'allow'. The interpretation of Krosnick & Schuman (1988) of the connotations explanation is correct: the answering scales have different meanings dependent of the question wording. The connotations of 'forbid' and 'allow' may be an explanation for this. But first of all it is feasible to check whether Schuman & Presser's claim that this might be "the sole source of the wording effect" is correct. And if not, which other explanations seem warranted. This was investigated in the meta-analysis that will be discussed in the next section.

II Explanations for the forbid/allow asymmetry. A meta-analysis

An affirmative ('yes') to 'forbid' does not mean the same as a 'no' to 'allow'. Schuman & Presser theorise that the sole reason for this may be the rather extreme connotations of both verbs. This would suggest that the asymmetry occurs in every question, and is about the same size all the time. The huge variances found in the first part of the meta-analysis (see section I.1) however, suggest that there is room for additional explanations. Therefore, in the second part of the meta-analysis it was explored which question characteristics, respondent characteristics or other characteristics determine variance in the size of the asymmetry. Insight into the 'causes' of the variance, may provide insight into the mechanisms that next to or in interplay with the connotations of the verbs explain the origin of the asymmetry.

Many explanations have been formulated in addition to the connotations hypothesis. One of these focuses on attitude strength, theorising that respondents with a weak or subtle opinion would be mainly responsible for the asymmetry. They find an affirmative answer too strong (because of the connotations of both verbs) and answer 'no', without realising that this implies an affirmative answer as well: 'no, not forbid' implies 'yes, allow' (Hippler & Schwarz, 1986; c.f. Fazio et al., 1982). This is the only explanation for the asymmetry that has been tested. Results are ambivalent however, probably partly because of differences in the operationalisation for the concept of attitude strength.

In addition many other explanatory hypotheses for variation in the size of the asymmetry have been offered. Most of these have not been tested. Next to the hypothesised influence of *psychological factors*, such as attitude strength, three categories of hypotheses can be distinguished. First there are *linguistic hypotheses*, which focus on the effects of indicators of linguistic complexity. Questions that are linguistically complex, would demand such a lot of working memory and attention in order to process the text, that respondents would not realise that not forbidding actually implies allowing. This would cause a greater asymmetry for linguistically complex questions (Schuman & Presser, 1981). Secondly, there are hypotheses that focus on the influence of *question content* on the nature of the asymmetry between 'forbid' and 'allow'. For example, if a question is about an issue that is forbidden at the time the question is posed, the verb 'allow' may get a more 'active' and 'changing' meaning than when the issue was allowed at the time the question was posed. Thirdly, characteristics of the *administration mode* of the question may influence the size of the asymmetry. For example: a question posed by phone or face-to-face gives the respondent less time to process the question and think of an answer, than a written questionnaire. The respondent has less time to realise the implications of a negative answer - therefore, the asymmetry found in oral questionnaires may by bigger.

Of course, it would have been possible to test all of these explanations for variation in the size of the asymmetry experimentally. However, most explanations were offered on the basis of results obtained with one question. So it seemed more useful to check first of all whether those explanatory suggestions are true altogether when analysing over all forbid/allow questions at the

[2] For technicalities of this design see Holleman (o.f.p.).

same time. This was done by coding in the original questions as many characteristics (like the ones mentioned above) as possible. These (15) explanatory variables were coded with either 0 or 1, so that the asymmetry size for, for example, oral questions (1) could be compared to the asymmetry size for written questions (0). The dependent variable was the difference between the answering percentage 'not forbid' and 'yes allow', again weighed for the number of respondents. For each of the coded characteristics three things were taken into account: significance of the effect, the amount of variance being explained (within and between experiments), and the mean effect size.

Results of this exploratory meta-analysis indicate that psychological factors (attitude strength) seem to be the most important explanation for variation in the size of the asymmetry. Linguistic complexity, content variables, or characteristics of the administration mode, seem of less importance. For example, issue complexity turned out to be an important indicator of asymmetry size, a result that supports hypotheses focusing on attitude strength. The more complex the issue of the question, the bigger the asymmetry. So mainly respondents with a weak or balanced opinion seem responsible for the asymmetry.

Also, results seem to indicate that the asymmetry is mainly caused by characteristics of the specific communicative task. This is a finding which is in line with the finding in the two experimental studies, that indicated the answers to forbid and allow questions reflect similar attitudes, expressed differently on the answering scale. Attitude questions in yes/no format force the respondents to localise (or form) the requested attitude in their heads and map their answers on to the answering scales. Especially 'no' seems a repository of different opinions, like "no, I do not think it should be forbidden/allowed", or "(no,) I have no opinion on this issue, so I do use the extreme yes-option", or even maybe "(no,) I do not agree with the presupposition in the question, so I do not answer affirmative".

III Conclusion

Since the 1940's a lot of research into the forbid/allow asymmetry has been carried out. Still there are a lot questions concerning the generalisibility of the wording effect, the exact nature of the asymmetry, and the causes of the asymmetry. In this paper some of these questions are investigated, bringing us closer to insight into the underlying mechanisms and causes of the asymmetry.

Research reported here demonstrates that the forbid/allow asymmetry exists all in all when analysis is done over all 52 forbid/allow questions. Variances are huge however, raising doubt as to whether the asymmetry is solely caused by the connotations of 'forbid' and 'allow'. The results of two correlational experiments show that the asymmetry occurs because the verbs 'forbid' and 'allow' cause the answering scales to get a different meaning. Forbid/allow questions measure similar attitudes, but the answers are expressed differently on the answering scales. An explorative meta-analysis indicates that attitude strength might be an important factor facilitating the occurence of the asymmetry. It may be the case that attitude strength interacts with the way 'true opinions' are put on the 'observed' answering scale.

References

Cicourel A.V. (1982), Interviews, surveys and the problem of ecological validity. In: *The American Sociologist 17*, 11-20

Fazio R.H., Sherman S.J., Herr P.M (1982), The feature-positive effect in the self-perception process: does not doing matter as much as doing? In: *Journal of Personality and Social Psychology 42 (3)*, 401-411

Hippler H.-J. & Schwarz N. (1986), Not forbidding isn't allowing: the cognitive basis of the forbid-allow asymmetry. In: *Public Opinion Quarterly 50*, 87-96

Holleman B.C. (offered for public.), The nature of the forbid/allow asymmetry: two correlational studies.

Jobe J., Mingay D.J. (1991), Cognition and survey measurement: history and overview. In: *Applied Cognitive Psychology 5*, 175-192

Jöreskog K.G. (1971), Statistical analysis of sets of congeneric tests. In: *Psychometrika 36*, 109-133

Krosnick J.A. & Schuman H. (1988), Attitude intensity, importance and certainty and susceptibility to response effects. In: *Journal of Personality and Social Psychology 54* (6), 940-952

Rugg D. (1941), Experiments in wording questions:II. In: Public Opinion Quarterly 5, 91-92

Schuman H. & Presser S. (1981), *Questions and answers in attitude surveys. Experiments on question form, wording and context.* Academic Press

Waterplas L., Billiet J., Loosveldt G. (1988), De verbieden versus niet toelaten asymmetrie. Een stabiel formuleringseffect in survey-onderzoek? [The forbid/allow asymmetry. A stable wording effect in survey research?] In: *Mens en Maatschappij 63* (4), 399-417

Kathrin Gieseking
FB 2: Computational Linguistics
University of Trier
D - 54286 Trier, Germany
E-mail: giesekin@ldv.uni-trier.de

TOPICAL PAPER - PSYCHOLINGUISTICS

Summary: The *Tuning Hypothesis* (TH) is a psycholinguistic model of human sentence processing. It predicts processing effort for ambiguous syntactic constructions based on their frequency in an individual's language input. This paper evaluates the predictive power of the TH as compared to non-frequency-based psycholinguistic processing models. It shows that while non-frequency-based models can predict processing effort for individual constructions correctly, the TH covers a wider range of phenomena with correct predictions.

Evaluating a frequency-based principle of human sentence processing

Introduction

In quantitative linguistics, frequency counts resulting from corpus analyses are often investigated under the focus of searching laws that are in force in language viewed as a system *(langue)*. However, quantitative methodology can also be applied in psycholinguistic research which tries to find laws that govern the *processing* of language within the human mind. A step towards this direction is made by the work reported in this paper. Here, too, frequency plays a crucial role. The basic research question investigated is:

> *In what way is the working of the human language processing mechanism influenced by the relative frequencies of linguistic units in its input?*

This question covers a vast field. It needs to be further specified in order to arrive at testable hypotheses. In the research reported here, only those aspects of processing will be considered that steer the *comprehension* of language, nothing will be said about language production. The linguistic units that are focused are ***syntactic units***, mainly on the phrase-level, and the working of the human language processing mechanism is considered only in as far as it governs the very first structural analysis of a sentence, the *initial parse*. The initial parse contains merely an analysis of the sentence that reveals the dependencies within the internal structure of the sentence. No deeper semantic processing is concerned.

It is uncontroversial that there does exist a correlation between frequency of linguistic items and ease of processing on the lexical level. E.g., psycholinguistic research has repeatedly shown that frequent words are processed faster than infrequent ones. On the syntactic level, however, the possible relation between the frequency of linguistic units and the speed of their processing has been very little researched so far.

A frequency-based approach for explaining human sentence processing

A first theoretical proposal concerning the influence of item frequency on language processing on the syntactic level *(sentence processing)* is made by the *Tuning Hypothesis* (TH) by Don Mitchell (Mitchell & Cuetos 1991; Mitchell, Cuetos, Corley & Brysbaert 1995). This psycholinguistic hypothesis claims that individuals keep internal statistics about received language input. These statistics about the *previous exposure* of an individual to language data

guide the processing of language and form the basis for decisions when the processing mechanism faces an ambiguous construction. It is continually updated by the analysis of the current input. The input data *tunes* the human language processing mechanism, with the result that frequent constructions are processed easier, i.e. faster, than infrequent ones.

Empirical evaluation of the TH

The ideal way for testing the TH would be to monitor the language input of an individual from birth on, to keep track of the frequencies of all kinds of syntactic units in the material and then subject the individual to psycholinguistic tests measuring his or her processing speed for diverse syntactic constructions. Comparing the frequency data and the processing performance in the experiments should yield either support or refusal of the TH. Obviously, this way of evaluation cannot be realized. Therefore, in an attempt to arrive at an approximation of the previous exposure of an individual, the frequency of syntactic units is counted in large amounts of language material. The resulting data is then compared to the results of psycholinguistic experiments which measure the processing time required by the subjects for these syntactic constructions. If syntactic constructions that are very frequent in the language material show faster processing times than infrequent ones, this can be viewed as supporting the TH. On the other hand, if frequency does not have an effect on processing time, this argues against the TH.

There are basically two on-line methods used in psycholinguistic experiments for measuring processing time for the initial parse. In *self-paced reading experiments* the experimental sentences are presented to the subjects word by word on a computer screen. The speed of the presentation is controlled by the subjects themselves. *Eye-tracking experiments* allow a more natural way of reading. A whole sentence is presented on a computer screen and it is measured how long a subject looks at each word of the sentence.
If an increase in processing time is measured using either of these methods, one can draw the conclusion that there is an increased processing effort connected with the experimental sentence, caused e.g. by unconscious processing problems or processes of reanalysis.
The material used in these experiments is typically structurally (at least temporarily) ambiguous material which allows the construction of minimal pairs, i.e. pairs of sentences that differ in only one parameter.

Corpus material

Since the claim of the TH to explain processing phenomena is not restricted to a single language, I will present empirical results from corpus studies in two languages, German and English. The German material is extracted from a corpus of three months of a German daily newspaper (taz - die tageszeitung) of 1993/94, the English material stems from the British National Corpus and contains a variety of text types. Each corpus contained about 70,000 sentences.

The type of structural ambiguity considered in this paper is the attachment ambiguity arising in constructions like (1).

(1) Manfred fesselte den Mann mit der Krawatte.
Manfred tied up the man with the necktie.
(1a) Manfred [fesselte [den Mann] mit der Krawatte]
(1b) Manfred [fesselte [den Mann mit der Krawatte]]

Here, the PP *mit der Krawatte* can be either attached to the verb *fesselte (high attachment/VP-attachment)*, carrying an instrumental function (1a), or to the noun *Mann (low attachment/NP-attachment)*, fulfilling an attributive function (1b).

Processing difficulties can usually be observed in structurally ambiguous constructions like (1) when the ambiguity is disambiguated by material that contradicts the subject's initial attachment choice. E.g., increased processing time should show up in the processing of sentences like (2) if the subject initially attached the PP to the VP and had to revise this attachment when the noun *Schnurrbart* appeared. Vice versa, sentences like (3) should produce increased processing time if the initial PP-attachment was to the NP and had to be revised to a VP-attachment.

(2) Manfred fesselte den Mann mit dem Schnurrbart.
... with the mustache.
(3) Manfred fesselte den Mann mit den Handschellen.
... with the handcuffs.

I carried out two investigations concerning this type of construction. I will present in each case the predictions for processing effort made by the TH and contrast them with the predictions made by other psycholinguistic models that claim predictive power with respect to processing effort but do not refer to frequency. Then I will compare the empirical results of psycholinguistic experiments investigating the relevant ambiguous construction with the predictions of the different processing models.

Corpus analysis I: Is there a correlation between the definiteness of the direct object NP and the attachment preference of the *with*-PP ?

The first investigation asks whether there exists a relation between the definiteness of the direct object NP and the attachment preference of the PP. The relevant constructions in this case are (4a) and (4b).

(4a) Manfred fesselte den Mann mit der Krawatte.
... the man ...
(4b) Manfred fesselte einen Mann mit der Krawatte.
... a man ...

A non-frequency-based psycholinguistic model capable of accounting for processing differences due to the definiteness of the article of the direct object NP is the model by Altmann & Steedman (1988), often referred to as *Referential Theory* (RT). The RT claims that modifier attachment ambiguities are resolved by recourse to higher-level contextual and referential information. This claim is specified in two processing principles.

(p1) *Principle of Referential Support*
An NP analysis which is referentially supported will be favoured over one that is not.
(p2) *Principle of Parsimony*
A reading which carries fewer unsupported presuppositions will be favoured over one that carries more.

In the case of isolated sentences, which are typical for most psycholinguistic experiments, (p1) cannot apply since no reference model is available to the reader. Therefore, (p2) is

applied in order to predict attachment preferences. This leads, according to Altmann & Steedman, to the postulation of a bias towards VP-attachment since an attributive NP-atttachment would presuppose at least two men, one of which needs to be specified further by a modifier, while the VP-attachment does not require the presupposition of at least two men and therefore is the more parsimonious reading.

The predictions that the TH makes for the mental processing of ambiguous constructions can be directly deduced from counts of these constructions and their respective readings in the corpus material. Counts in the English corpus material led to the following distribution:

	VP-attachment	NP-attachment
def. NP	75 (95%)	4 (5%)
indef. NP	22 (39%)	34 (61%)

It follows from these data that the TH makes the same predictions as the RT for sentences containing definite NPs, i.e. a VP-attachment preference is predicted. In addition, the TH predicts an NP-attachment preference for sentences with indefinite direct object NPs. Note that the RT does not make explicit predictions for indefinite NPs.

Spivey-Knowlton & Sedivy (1995, Experiment 5) have done a self-paced reading experiment using sentences containing the relevant attachment ambiguity, varying the definiteness of the direct object NP. In this experiment, sentences containing definite NPs led to shorter reading times if the PP was attached to the VP, while sentences containing indefinite NPs led to faster processing if the PP was attached to the NP.
These empirical results are compatible both with the predictions of the RT and the predictions of the TH generated by the corpus counts.

Corpus Analysis II: Is there a correlation between the verb position and the attachment preference of the *mit*-PP?
The second corpus study, done on German material, deals with the question whether the position of the verb is correlated with the attachment preference of the PP. In German, both verb-second (5a) and verb-final (5b, 5c) sentences are grammatical.

(5a) Manfred fesselte den Mann mit der Krawatte.
(5b) Manfred hat den Mann mit der Krawatte gefesselt.
Manfred has the man with the necktie tied up.
(5c) ...daß Manfred den Mann mit der Krawatte gefesselt hat.
...that Manfred the man with the necktie tied up has.

A psycholinguistic model that makes predictions about the influence of the verb position on the PP-attachment ambiguity is the Parameterized Head Attachment principle (Konieczny, Scheepers, Hemforth & Strube, 1994; Konieczny, Hemforth, Scheepers, & Strube, in press). It proposes that the initial syntactic analysis is determined by applying three hierarchically organized sub-principles (p3-p5):

(p3) *Head Attachment*
If possible, attach a constituent to a phrase whose lexical head has already been encountered.

If further attachment possibilities exist for the critical constituent, then

(p4) *Preferred Role Attachment*
Attach Constituent *i* to a phrase within the current clause whose head highlights a theta-role for *i*.

If further attachment possibilities exist for the constituent, then

(p5) *Recent Head Attachment*
Attach the constituent to the phrase whose head was encountered most recently.

In order to arrive at a prediction for an ambiguous construction these principles have to be applied in the specified order. Generating processing predictions for (5a), (p3) is not applicable since both potential heads *(fesselte, Mann)* have already been encountered. If none of the potential heads licences a theta-role, (p4) cannot be applied either, and the application of (p5) leads to the prediction of an NP-attachment preference. If one of the potential heads does licence a theta-role, an attachment preference to this head (usually the verb) is predicted according to (p4).
Generating predictions for (5b) and (5c) requires only the application of (p3). Only one potential head, the direct object NP, is available at the point of ambiguity. Therefore, a PP-attachment preference to this NP is predicted.

The corpus results for these types of constructions in the German material show the following distribution:

	VP-attachment	NP-attachment
Verb-second, verb does not licence a *mit*-PP	11 (29%)	27 (71%)
Verb-second, verb licences a *mit*-PP	7 (87,5%)	1 (12,5%)
Verb-final	37 (39%)	59 (61%)

Consequently, the TH makes the same predictions for processing effort as the PHA principle: A VP-attachment preference is predicted for sentences with verb-second position and verbs that licence a *mit*-PP. For the other two cases, an NP-attachment preference is predicted.

The empirical reading time data gained in an eye-tracking experiment by Konieczny, Hemforth, Scheepers & Strube (in press) supports the predictions of both models. Reading times were shorter for NP-attachments except in the case of sentences with verb-second position containing verbs that were biased towards taking a *mit*-PP as an argument.

Please note that this data contradicts the predictions of the well-known garden path-model by Frazier (1987), which claims that VP-attachment is favored independent of the valency of the verb because of the minimal attachment principle.

Conclusions

In both investigations described in this paper the TH was able to arrive at predictions of processing effort that were confirmed by the empirical reading time data.
Specifically, the TH was able to predict the results for a wider range of phenomena than the non-quantitative approaches presented here: The RT with its principle of referential support and principle of parsimony cannot generate predictions about the relation between verb position and PP-attachment preference, while the PHA principle cannot predict a relation between definiteness and PP-attachment preference. The TH can account for both kinds of data relying on a single principle, frequency, and therefore must be viewed as the most parsimonious approach.
While the results presentend in this paper are not sufficient yet for confirming the validity of the TH, they certainly indicate the existence of a correlation between frequency and processing effort not only on the lexical level, but also on the level of syntax.

References

Altmann, G. & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*, 191-238.

Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart (Ed.), *Attention and Performance XII: The psychology of reading* (pp. 559-586). London: Lawrence Erlbaum.

Konieczny, L., Hemforth, B., Scheepers, C. & Strube, G. (in press). The role of lexical heads in parsing: Evidence from German. *Language and Cognitive Processes*.

Konieczny, L., Scheepers, C., Hemforth, B. & Strube, G. (1994). Semantikorientierte Syntaxverarbeitung. In S. W. Felix, C. Habel, & G. Rickheit (Eds.), *Kognitive Linguistik: Repräsentation unf Prozesse* (pp. 129-158). Opladen: Westdeutscher Verlag.

Mitchell, D. C. & Cuetos, F. (1991). The origins of parsing strategies. In C. Smith (Ed.), *Current issues in natural language processing*. University of Austin, TX: Center for Cognitive Science.

Mitchell, D. C., Cuetos, F., Corley, M. M. B. & Brysbaert, M. (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records. *Journal of Psycholinguistic Research*, *24*, 469-488.

Spivey-Knowlton, M. & Sedivy, J. C. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition*, *55*, 227-267.

Quantitative Analysis of Different Processing Patterns in Listening Comprehension by L2 Listeners

Setsuko Wakabayashi (Miss)
Himeji Dokkyo University
7-2-1 Kamiohno, Himeji, 670 Japan
e-mail: setsuko@himeji-du.ac.jp

Introduction

L1 (mother tongue) listeners normally share the same language background with speakers as a basic communication tool, and are able to process, on line, a sufficient number of information units, decoding sounds into meanings in the comprehension process. Conversely, L2 listeners are often unable to process enough information units, since they have less command of L2. Utilizing contexts and general knowledge relating to pieces of information extracted from the input, L2 listeners might have to make specific efforts to compensate for their shortcomings in order to interpret the meaning of messages.

Individual computers are provided for listeners to have listening practice at their own pace and with random access to the materials in order to trace their natural listening comprehension patterns. The study history of about 60 university students of L2 English was collected automatically for analysis.

Analysis of study histories shows that among L2 listeners information units were chosen and processed in different ways, exhibiting different processing patterns. In one respect processing was fairly similar with L1 processing: Information units of larger sizes and/or at higher levels were dominantly used. Students were found to work repeatedly through small information units when they were trying to catch the sound; but once they tried to interpret the meaning of small units, they required larger units. L2 listening requires more elaborate processes and the processing patterns can be categorized into three types.

General concepts of language processing

Greene (1986) argues for heterarchically controlled processing in which information from different types of processing can be pooled before deciding on appropriate representation for linguistic inputs.

In cognitive psychology and psycholinguistics, knowledge is considered to be central for processing and is utilized in *top-down* inference (Lackman, Lackman and Butterfield: 1979 and Abe, Momouchi, Kaneko and Yi: 1994). Although Winograd (1983) favours the bottom-up process where sounds are sequentially processed to reach meaning representation, he also supports the view of knowledge as a basis of processing.
Abe et al. (1994) explain language processing by means of situational meanings utilizing linguistic and general knowledge. They claim that interpretation of speech is achieved when listeners cognitively form a semantic representation of speech and a speaker's world

expressed in a certain situation and a context. Sperber and Wilson (1986) also emphasize the important role of relevance to context and meaning in processing.

L2 listeners' comprehension processes also utilize context and general knowledge. It is noticeable that L2 listeners with less language competence require higher levels of information rather than lower levels in order to have better reference to their general knowledge or to contexts and to achieve their interpretation.

Data collection

Method

Discrete point questions for vocabulary meanings and integrative questions for the text understanding (Davies: 1990:34f) were provided. Two different conditions were tested: Audio (only) and audio video presentation to see whether there would be changes in the way they were processed. No written text information was given unless the video text itself exhibited some (e.g. small writing in the background). For the discrete point questions, students had to provide Japanese definitions or translations of particular English words in the text. By clicking on numbered question buttons the students could, at any time, hear a recording of any of these test words. The integrative questions were given by written text in Japanese.

Listeners were allowed to spend about one hour to answer questions. The order in which the modes (audio only or audio video) were presented was controlled by the author. Approximately half received audio only first and the others had audio and videos as the first condition. The same questions were repeated under the second condition. Otherwise where and how to listen to the text were left to the listeners. Study history (e.g. which audio units, for how long, in which order, and their other choices) was automatically collected and stored on the hard disk of the computers for later analysis. At the end of each session students were invited to write comments on any aspects of the task which they had found problematic and provide any other comments.

Subjects:

About 30 lower intermediate learners of English at a Japanese University (1st year, English major) from two different classes (total 60 students) participated in the experimental classes for two weeks. About one hour was spent per session containing both conditions and each student participated in two sessions.

Material:

A distance-learning video text (filmed and produced by the Media Centre at the University of the South Pacific (USP)) GE101 Introduction to Physical Geography, by Professor Patrick Nunn was digitized and edited into 2-3 minute selections, each selection being the materials for one weekly session.

Each digitised audio-visual text was prepared so that students could freely have access to

it via four different sizes and levels of units:

the largest size (e.g. paragraphs) for highest level [topics],
large size (e.g. paragraph) for high level [major parts of topics],
medium size (e.g. sentence) for middle level [parts of the major parts]
and small size (e.g. words and phrases) for low level [units of information for parts at middle level].

An example stackcard of each type of sizes and levels of information and of the screen as seen by the students is attached (See Appendix).

Mechanism:

Digitized visual-audio text (with access at will to pause, repeat, stop, and play functions, and to speed control) was provided. One click of any button allowed listeners to receive instantly whatever they have chosen. They could choose to play any size of unit, in whichever order, and whenever they wanted it. The chosen units continued to play unless listeners take any further actions.

Results

1. Features found in the study history

1-1. Different listening patterns

Subjects sometimes changed their chosen size of information units from larger to smaller, and from smaller to larger. They also often stayed in the same size and/or levels. Listening patterns can be categorized into three groups by the listeners' dominant use of the different sizes of information units and the time they spent with a chosen unit: 1) One pattern is characterized by the usage of large units with less repetition: That is the processing of large units of information, rather than repeated access to small units. 2) A second pattern is the usage of small units with many repetition; that is processing smaller units of information. 3) The third pattern is the usage of all sizes of information.

1-2. Mean processing size

The study history shows that subjects most frequently chose to play larger information units. The mean number of processing units chosen by listeners, for example, in one of the tasks was 6.4 (small information units), which places close to large size (largest size 10.6, large size 4.8, medium size 2.5 and small size 1). Listeners started from larger sizes rather than small even though they could have free access to any units. Replay also tented to be started with larger information units.

2. Features found by analyses

2-1. Access to contexts and general knowledge

Analysis seems to indicate that listeners required larger amounts of information for L2 listeners' processing in order to have effective access to contexts and their general knowledge. They appeared to utilize contexts and general knowledge in compensation for their difficulties with small processing units due to their limited L2 knowledge. This

was the dominant strategy for inferring meanings.

2-2. Distinguishing L2 sounds
Most students reported that they used small units rather than large ones when they were trying to clarify the sounds of the units in their attempts to refer to previous knowledge.

Discussion
1. Higher levels of information were dominantly used, which closely relates to general knowledge. The shortage of linguistic knowledge seems to impel L2 listeners to utilize context their general knowledge. Listeners managed to grasp the gist even out of parts of messages that they understood imperfectly. Rost (1990) comments that much mishearing goes undetected, or is self-corrected by the listener, through the use of contextual information. A listeners interpretation does not have to utilize every single unit to construct an inference and does not have to be exactly the same as the speaker's intent. The study of auditory perception also reinforces the claim that subjects do not listen to what they are precisely and physically hearing, but utilize what they have previously processed (Otake: 1995). L2 listeners exhibited difficulties comprehending small units of information and thereby relied on larger information units as well as context and their general knowledge and contexts for top-down inference.

Other studies also support the importance of the previous knowledge which listeners have. The results are as follows: 1) previously known information proved to be effective for comprehension of new information (Haviland and Clark: 1974), and 2) that listeners are more likely to be able to comprehend and remember a passage if they know what it is about prior to their listening (Bransford and Johnson: 1972). Yamadori's claim (1985) too supports the significance of previous understanding in language processing. As L2 listeners find it easier to decipher when they understand the meaning, dominant employment of larger information units in L2 listening is understandable.

2. L2 listeners seem to require two stages of inference: meaning for information units and the meaning for the message. Processing units have to be larger for inferenceing meanings, but individual preference seems to influence the choices of processing and to exhibit different processing patterns. For example, some subjects tend to do discrete point questions first and others do them as their listening process proceeds. Various listening patterns also show that some subjects use smaller units and listen to them repeatedly, some use rather larger units and go through them few times and some use a mixture of both types. At present the different patterns indicate that individual preference is a main cause for them rather than choices of sizes and levels of information units available for various processes.

3. It is not necessarily lower language levels of performers who employ lower levels of information and build up small bits of information into larger bits in order to complete their interpretation. What seems clear is that listeners require larger units of information

in order to get meaning and smaller units of information to discriminate sound. But when listeners have to process large amount of information at a time they seem to have short-term memory problems. Some of them repeatedly utilize smaller information or write down the meaning of the units and construct the meaning gradually. Also highlighted is that some listeners conduct processing step by step, while others use spiral processing.

Conclusion
Language capacity limits the sizes of information units that can be coped with in processing but this study indicates that L2 listeners employ various sizes of information unit in their individual performance. It is left for future researches to investigate what causes those different processing patterns, but it is clear in this analysis that different patterns exist among L2 listeners.

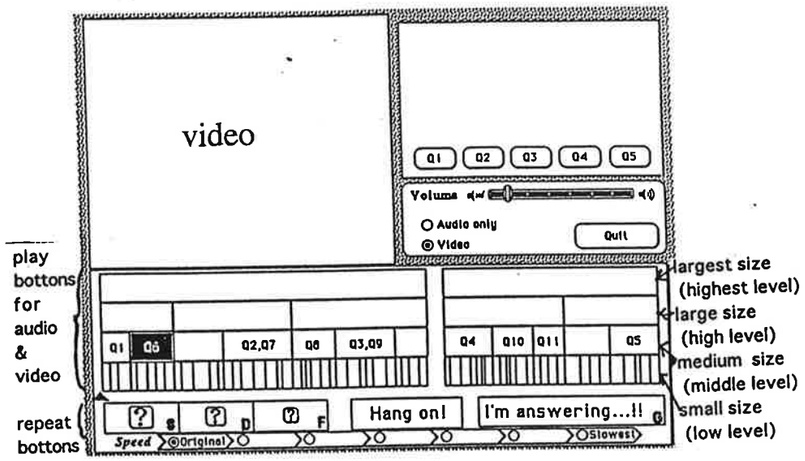
Acknowledgments
This research was supported in part by the Information Science Centre of Himeji Dokkyo University with special research funds granted by Himeji Dokkyo University. I appreciate my students' diligent performance in class, which naturally offered good data for the study.

I received support and comments from my research colleagues, Jun-ya Morishita, and Koichi Sonoda. The design and programming of the HyperCard Stacks were undertaken by my research colleague, Tomoyuki Sano, to whom I am indebted. I would also like to acknowledge valuable comments from Patrick Griffiths of USP, Fiji and Takeshi Kinoshita, Tokyo University. I owe enormous gratitude to Koichi Kurahashi for invaluable support, comments and encouragement. I thank USP and Professor Patrick Nunn for permission to use the GE101 video material.

References
Abe, J., Momouchi, Y., Kaneko, Y. and I, K (1994) *Human Language Information Processing* in Cognitive Science & Information Processing - 12 (in Japanese), Tokyo: Science Sha.
Bransford J.D., and Johnson, M.K. (1972) 'Contextual prerequisites for understanding: some investigations of comprehension and recall', *Journal of Verbal Learning and Verbal Behavior*, 11, pp. 717-26.
Davis, A. (1990) Principles of Language Testing. Oxford: Basil Blackwell.
Greene, J (1986) Language Understanding: A Cognitive Approach, Milton Keynes: The Open University Press
Haviland, S.E. and Clark, H.H. (1974) 'What's new?' Acquiring new information as a process in comprehension', *Journal of Verbal Learning and Verbal Behavior*, 13, pp. 512-21.
Lackman, R., Lackman, J.L. and Butterfield, E.C. (1979) *Cognitive Psychology and Information Processing: An Introduction* in Cognitive Science & Information Processing - 6 (in Japanese), Tokyo: Science Sha.
Otake, T. (1995) 'Auditory Perception' in Y. Otsu (Ed.) *Psycholinguistics* (in Japanese), Cognitive Psychology 3, Tokyo: University Tokyo Press.

Rost, M. (1990) Listening in Language Learning, London: Longman.
Sperber, D. and Wilson, D. (1986) Relevance: communication and cognition. Cambridge, MA: Harvard University Press.
Winograd, T. (1983) Language as a cognitive process. Volume 1: *Syntax*. Reading, MA: Addison-Wesley.
Yamadori, A. (1985) The Mind Observed through the Brain (Nou kara mita Kokoro) (in Japanese), Tokyo: NHK Broadcast Inc.

Appendix



a short summary:
L2 listeners usually have to comprehend within their limit of linguistic competence. Quantitative analysis was carried out in order to find 1) types of processing with various sizes and levels of information, and 2) how such variable types of processing emerge. The analysis indicates that different types of processing patterns are categorized by the choice of various information units, and that individuals use them in their own ways.

key words:
listening comprehension, processing patterns, information units.

a specification of the topic area:
psychолinguistics

Quantitative Aspects of Finnish Wernicke Speakers' Narratives

Jussi Niemi

Linguistics
University of Joensuu
FIN-80101 Joensuu, Finland
jussi.niemi@joensuu.fi

Topical paper
Psycholinguistics

Aphasic language deficits are typically gradient deviations from the nonpathological norms (*graceful degradation* in Parallel Distributed Processing terms). The present study concentrates on pathological syntactic patterns in Wernicke's aphasia in a grammatically free word-order language (Finnish). In addition to evincing a dissociation between phonology and syntax on one hand and lexical semantics and pragmatics on the other, the aphasics show, inter alia, decreased syntactic complexity, operationalized through various structures, in their free output.

The notions of external evidence, psychological reality and learnability have been used by autonomous linguists to constrain the grammatical descriptions since the late 1960s. The data-sources have included, inter alia, language acquisition patterns, contact languages, language decay and death as well as language games and pathologies, most notably adult aphasia. Thus it comes as no surprise that the establishment of psycholinguistics as a valid field of scientific enterprise also took place during that decade. Although psycholinguists have developed and adopted many interesting on-line techniques to analyze how normal language-users process and represent language in space and time, the off-line output of the damaged adult language processor (aphasic mind) still serves as a highly useful window into the intact language processor.

In the 1970s the psycholinguistic approach towards aphasic language emphasized the juxtaposition of two linguistic explanations of Broca's and Wernicke's aphasia. At one extreme, Broca's aphasia was characterized as being 'lexicon without syntax'. The proponents of this view argued for a syntactic deficit as the ultimate linguistic cause of so-called Broca's agrammatism (at least in languages like English). At the other end of the syntax vs. lexicon spectrum there was a view stating that Wernicke speakers show a lexical-semantic deficit coupled with unimpaired syntax (Wernicke's aphasia being 'syntax without lexicon'). Hence the internal conceptual partition (modularity) of grammar received a corresponding division in the localization of language functions and modularity in processing linguistic information.

The specific aim of the present study is to analyze the syntactic output structures of two Wernicke speakers in Finnish, a richly-inflected language with relatively free

grammatically-specified word-order. The analysis procedures of the present study are a modification of a quantitative analysis carried on normal Finnish by Hakulinen, Karlsson and Vilkuna (1980). The quantitative analysis of Hakulinen et alii was modified for the present pathological purposes by coding each clause for 46 structural variables chosen to represent a wide variety of grammatical patterns, ranging from inflectional morphology through syntax to information structure.

The following findings are, inter alia, hard to account for using any lexical account of Wernicke's syntactic aberrations: The Wernicke speakers have a low number of complex syntactic constituents, and their word order and case marking show an overuse of the canonical patterns of the language. In the oral presentation I will discuss these and other significant observations of the aphasics as well as the intriguing disassociation between phonological and syntactic aberrations, on one hand, and lexical semantic and pragmatic, on the other.

REFERENCE

Hakulinen, A., Karlsson, F. & Vilkuna, M. 1980. *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus*. Publications of the Department of General Linguistics, 6. Helsinki: University of Helsinki.

TRAJECTORIES OF FRACTAL ATTRACTORS IN PREFERABILITY SITUATIONS

Vladimir Otlygin, Moscow

The paper deals with the study of the brain activity potentials by normal subjects and those with psychopathological deceases (e.g. schizophrenia) in preferability situations - i.e. those of choosing preferable concept out of two or more given concepts (e.g. in answering questions, such as what is more important for you - your family or your job, your children or your health, to what extent is X more important to you than Y, etc.).

It is supposed that solutions in such situations, as well as their verbal realization depend on systemic mechanisms, and that the preferability situations build a dynamic system with unknown parameters

Using (1) the Bocher, Timsit-Berthier, et al. (1990) method of registration of the conditioned-negativ brain potentials (CNBP) build by the delayed answers in the preferability situations, and (2) the method of the EEG-analysis suggested by Galez, Bablojanz (1991), trajectories of fractal attractors for each dual preferability situation (situation with two answers) have been obtained. It is shown that these trajectories are:

(a) coherent in the preferability situations with structurally complete sentence (containing subject, predicate and object) and incoherent (discontinuous on the set of attractors) when the preferability situation is expressed in anaphorical from, as denotat or as an close denotats.

(b) stable within a given group of subjects (normal subjects, subjects with various psychopathological deviations / diceases). In particular, for subjects in the depressive state the trajectory is cut into some pieces. The latter phenomenon makes it possible to use the method described above in the diagnostics of psychopatological deviations.

Based on the representation of the preferability situations on Mandelbrots fractal sets (Paitgen, Richter 1993), it is demonstrated that the trajectory of the fractal attractors in the preferability situations with two and with more than two possible answers differ from each other.

References:

1. Bocher, K.B., Timsit-Berthier, M. Schoenen, J. et al. (1990) Headache, 30, 604-609
2. Pajtgen, C., Richter, P. (1993) Krasota fraktalov (The beauty of fractals) Moscow: Mir [in Russian]