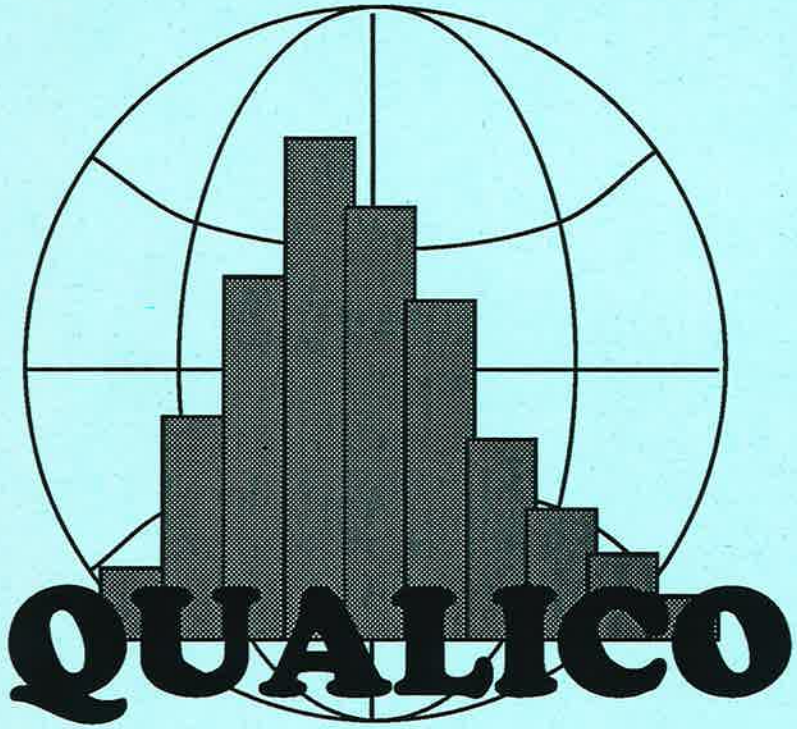


ABSTRACTS

QUALICO 91

September 23-27, 1991

University of Trier, Germany

CONTENTS

G. Altmann

SCIENCE AND LINGUISTICS

Science is a triple that can be described using the symbolization of Bunge as

Science = <Object, Approach, Theory>.

All of these components must be present in research, the problematic one in linguistics is theory. Objects form a set from which one can choose freely, i.e.

Object = {Thing, Property, Relation, Structure, Function, Process, History, System}.

The analysis of systems is the most complex task in linguistics whereby all other objects are involved.
The approach is again a structure defined as

Approach = <Aspect, Aim, Problem, Method>.

Though aspect concerns the ontology of the object, it is merely our view of it. We can consider it from evolutionary, structuralistic, generativistic, emergentistic, atomistic, systemic, synergetic, holistic, historical, static or dynamic, deterministic or stochastic etc. points of view. The aspect does not say us what the object is, it is our conceptual construction.

Aim = {Application, Description, Classification, Prediction, Understanding, Explanation}

Linguistics concerned with the description of rules, the so called "standard linguistics" is applied linguistics even if its conceptual apparatus is sometimes very complex. In quantitative linguistics, as in all riper sciences, explanation is the highest aim but it can be reached only if we can embed a linguistic phenomenon into a discernable nomological pattern (W. C. Salmon). To this end we must have a set of laws (not rules) or self-regulation schemes like that of R. Köhler yielding functional explanation through a lawful concatenation of phenomena.
Problems arise almost automatically if we have an aspect and an aim.

Methods = <Procedures, Conceptual means>

The procedures are well known. They are e.g. observation, measurement, introspection, experiment, induction, deduction etc. The conceptual means are either qualitative or quantitative. Qualitative means are not only verbal descriptions but also logic, set theory, algebra, graph theory etc. while the quantitative ones are combinatorics, probability and statistics, difference and differential equations etc. The old problem: "we need qualities, not numbers" is a misunderstanding. There are neither qualities nor quantities in the nature. These are merely properties of our concepts. What we need are laws.

Theory = <Concepts, Conventions, Hypotheses>

The existence of concepts is a necessary but not a sufficient condition for a theory though many "theoretical linguists" merely explicate concepts.
Conventions like definitions, operations, criteria etc. must be present in all sciences. They are artificial constructions not inherent in data.

M. Bunge distinguishes four kinds of hypotheses: Guessings that are neither derived nor tested; inductive hypothesis arising as empirical generalizations of observations, e.g. language universals, grammatical rules, sound "laws" etc.; deductive hypothesis derived from a theory but not yet tested empirically; laws which are statements derived from a theory, from axioms or other laws, tested and sufficiently corroborated empirically, holding for all languages in all times. A theory must contain at least some laws though it always contains many inductive hypothesis. Scientific hypothesis must be testable and improvable. Laws are mostly statements about hidden mechanisms that give rise to linguistic phenomena. A consequence of a law can be compared with linguistic observations. Theoretical linguistics is a science investigating processes, structures, functions and self-regulation of systems, and setting up systems of laws explaining language phenomena.

M. V. Arapov

WORD VALUE AND VOCABULARY STRATIFICATION

1. The purpose. The paper deals with the problem of stratification of the word-stock. A method is suggested for singling out elements which create its "upper layer", its basic or core vocabulary comprising the most valuable - from an aspect - words. When deciding whether a given word belongs to the core vocabulary or not, two macroscopic indices are usually taken into account - the overall frequency of the word in a sample and the number of texts in which the word occurs (or an appropriate function of one or both of the arguments). The criterion proposed in this paper is based rather on micro- than macro characteristics of word: a word is fit for the core vocabulary if comparison reveals a high degree of stability of using the same word in different texts.

2. The underlying idea. Each word is a focus of some shared extralinguistical experience (or knowledge) either originally expressed in the present text by means of a given word or gathered previously, and merely reproduced in the text. Usage of identical words sets up a relation between texts transforming their conglomerate into a dynamic net structure - social memory. The higher the value of experience stored in the memory, the higher is the degree of stability of expressing that experience in networked texts. That is, the core of word-stock consists of items whose usage in one text can be easily predicted on a basis of their usage in another. In an oblique way, measures of variability (such as ratio of texts in sample that comprise the word, variance of its frequency etc.) also characterize the degree of stability, but the straightforward technique to estimate that degree is by comparing glossaries of texts itselves.

3. The suggested process of selecting words includes two stages: first, the distance between a pair of compared texts (that is, between their glossaries) is measured and the obtained results are used to standardize data for separate words; second, in compared glossaries, the words, the differences of which in usage - measured on a standard scale - are found being less then a chosen threshold are selected for the core vocabulary. There are two species of words that interfere with the measurement of the distance and must be eliminated from glossaries beforehand. The first class consists of words born and mostly vanished with an individual text. Experience concentrated in these words is not actually nation-wide shared, it can be interpreted in the framework of this text only. The short span of life results in their place in the frequency dictionary and prompts methods of their elimination [1]. The second class includes all syntactic words. The knowledge concentrated in them is shared but it is strictly linguistical one. However, this knowledge is indispensible for the representation of the rest of human experience.

4. The distance between texts. The two classes having been excluded glossaries, for the rest of words the value

$$R(x) = \log i(x) - \log j(x),$$

where i(x) and j(x) are ranks of word x in compared texts, had

a normal (Gaussian) distribution N(m,s) with mean m and standard variation s, as it was supposed in [2]. The distribution is actually a one-parameter one because under the assumption that the correspondence between glossaries is symmetrical, parameters of location m and scale s are functionally related. Each of them can be used as a measure of resemblance between the texts. Unlike lexical items bearing extralinguistic experience, normally distributed statics for the syntactic words y will be

$$Q(y) = i(y) - j(y).$$

That is, the invariant characteristic of usage for a full word is a logarithm of ratio of its frequencies (or ranks) in the networked texts whereas the corresponding characteristic for a connecting word is a difference of such frequencies.

5. The word value measurement. To standardize the characteristic of usage R(x) and make it independent of the frequency of the word x and the resemblance of texts, let us transform it as follows,

$$T(x) = \text{abs } [R(x) - m] / s$$

For instance, a word x can be considered as pertaining to the core vocabulary if T(x) is lower then some chosen threshold for any pair of networked texts. The suggested technique is too tedious to be widely used unless in a simplified variant, but it may be helpful to elucidate what is a basic vocabulary.

REFERENCES

Arapov, M.V., M.M. Cherc (1983): Mathematische Methoden in der historischen Linguistik.

Arapov, M.V. (1988): Kvantitativnaja lingvistika.

Paraphrasing before Translation: On Machine Translatability and Sentence Complexity

Chao-Huang Chang* and Gilbert K. Krulee
Department of EECS, Northwestern University
e-mail: changch@eecs.nwu.edu

Extended Abstract

1 Introduction

Why is the development of a practical machine translation system (MT) such a difficult task? The answer is perhaps that we expect too much. In this paper, we are proposing a new model of machine translation based on the concepts of a sublanguage and of linguistic complexity. First, the sentences of the input are classified into four categories in terms of ease of translatability from the most to the least direct. Second, the translation complexity is defined as the maximum of three components: lexical, syntactic, and semantic complexity. Accordingly, a rule-based *Paraphraser* is used to convert a class of sentences, machine-translatable after paraphrasing, into the class of machine-translatable. It can be considered as a process of reducing translation complexity, converting the deviations from the standard language, or handling the periphery as in the universal grammar formalism. We suggest that to relieve the load of pre-editor, traditional machine translation system, and post-editor, every machine translation system should include a paraphraser component.

2 Sublanguages

The concept of working within a sublanguage has considerable relevance for the development of MT. Creating a MT system for a given sublanguage is much more practical than for the full set of possibilities for a given language. This important concept has been applied to research in text processing by Sager [1972]; Kittredge and Lehrberger [1982]; and Grishman and Lehrberger [1986].

Lehrberger [1986] divides a sublanguage L' of a natural language L into two parts: restrictions on and deviations from the grammar of the standard language L_{std}. According to the assumption, any construct in a sublanguage that deviates from L_{std} can be paraphrased into L_{std}. He also mentioned the difficulty of formalizing the paraphrasing in a principled way.

3 The Paraphrase-before-Translation Model

Based on the concept of sublanguage [Lehrberger, 1986], we are proposing a new translation approach as follows: The translating process is divided into 3 stages. First, *paraphrase* the input sentence in the sublanguage S' of the source language into the standard source language S_{std}. The paraphrasing may be an iterative process, reducing the lexical, syntactic, and/or semantic complexity of the sentence. Then, a normal *translating* module (for example, those systems in [Nirenberg, 1987, Slocum, 1987]) translates the paraphrased sentence to standard target language T_{std}. Finally, an optional *refinement* module can be used to generate the corresponding sublanguage T' of the target language.

As an example, if we are developing a Chinese-to-English MT system to translate Chinese News Headlines into English (an undergoing experiment), the sublanguage is first paraphrased into standard Chinese. Then, the paraphrase is translated into standard English. Finally, an optional module can be used to refine the version in standard English into the corresponding sublanguage, i.e., English News Headlines.

Note that the so-called standard languages are not necessarily well-defined. For practical purposes, we can take any existing MT system and let the source language sentences it can translate acceptably be the standard source language and the corresponding output target language sentences the standard target language. In other words, the parsing grammar and generation grammar rules in the selected MT system define the standard source and target languages. The next step in the development of this model is to classify sentences into four classes according to their translatability as follows.

Class I: Machine Translatable without Paraphrasing

The machine translation systems currently available [Slocum, 1987] can be used to translate this class of sentences successfully. Thus, Class I doesn't include the (linguistically) difficult sentences with unresolvable (lexical or structural) ambiguities, ellipsis, anaphora, metaphor, and so on. The output sentences from computers should be (syntactically and semantically) accurate and intelligible. For these sentences there is a direct correspondence between the syntactic forms used in both source and target languages.

Class II: Machine Translatable after Paraphrasing

The proposed PBT model is designed for handling Class II sentences. A sentence is first paraphrased into an equivalent Class I sentence. After that, we can simply use a currently available system to translate it. To achieve the task, we need to develop a Paraphraser and a Paraphrasing Rulebase.

*This research has been supported by the Industrial Technology Research Institute, Taiwan under a grant for doctoral study to the author. The address for correspondence with the author is: 1915, Maple Avenue, #419-1, Evanston, IL 60201, U.S.A.

Class III: Human Translatable

From the viewpoint of a translator, all sentences should be translatable. Probably, in some sublanguages such as Weather Reports, the statement is true. However, because of the linguistic distance between Source Language(SL) and Target Language(TL), it is argued that some sentences are not translatable. See the description for Class IV.

Although the sentences in this class can be translated by humans, they are not expected to be successfully translated by computers in the near future.

Class IV: Not Translatable

If SL and TL are in different cultural backgrounds(such as Chinese and English), some features in the SL will be almost impossible to translate into TL. As Bloom [1981] indicated, an English Counterfactual sentence is very hard to translate accurately into Chinese since the logics of these two cultures are very different. Literary works such as poems also fall into this class.

Translation Complexities

A natural language sentence has its linguistic complexity(lexical, syntactic, and semantic), just as a computer algorithm has its computational complexity(time or space complexity). For a translator, some sentences are much easier to translate than the other. Thus, a sentence can be assigned a "Translation Complexity" depending on the difficulty a translator will encounter during translation.

For convenience of discussion, we divide the linguistic complexity for translation into three parts: lexical, syntactic, and semantic. We should also include pragmatic(context) complexity.

The sentence complexity is defined as follows:

$$TranX = Max(LexX, SynX, SemX).$$

where TranX is the translation complexity for a sentence, LexX stands for Lexical Complexity, SynX Syntactic Complexity, and SemX Semantic Complexity.

Lexical Complexity

A word can be easily translated if it has only one fixed translation in TL. For example, proper names or terminologies in a special domain usually have unique (or standard) translation. Another example is idioms and other items in a phrasal lexicon. (See [Becker, 1975].)

However, most common words are with multiple meanings (polysemy) or at least with multiple translations. For example, English verbs "take" and "make" can be translated into a number of Chinese words with different meanings. Similarly, the Chinese verb (also a noun) da can be used to mean "beat", "buy", "thunder", "dozen" and so on (more than 100 translations depending on context, even in a concise Chinese-English Dictionary [Liang, 1987]).

Thus, we can classify the SL words according to the mapping to the TL words as follows:

one-to-one mapping Proper nouns, terminologies, idioms, etc.

N-to-one mapping The TL vocabulary is a subset of SL one. It is no problem to MT systems except, perhaps, taking care of the problem of repeating words.

one-to-N mapping Lexically or Categorial ambiguous words, Superset words(from SL to TL) and other multiple-translation-words. One common superset word from Chinese to English is the pronoun ta meaning "he", "him", "she", or "her".

N-to-N mapping Composition of the two mappings above.

one-to-zero mapping No corresponding words in TL.

zero-to-one mapping Corresponding to some implicit semantic content.

Syntactic Complexity

If two languages have similar or parallel syntactical structure, the translator can almost mechanically map SL sentences into TL ones, even in different word-order. Thus, one way to define the syntactic complexity is using the syntactic similarity. Unfortunately, sentences can be very complex (nested structures, embedded clauses, multiple attachments, ...) but still be simple in the syntactic complexity from a translator's point of view if there are corresponding syntactic structures between SL and TL.

Therefore, the syntactic complexity is defined differently between translation and comprehension (computer or people). In a sense, the syntactic complexity is absolute for text comprehension and relative (to TL) for translation. Of course, these two are correlated but not necessarily proportional. Usually, a sentence difficult to comprehend is also difficult to translate. It is expected that we can borrow some linguistic complexity theory for translation from those for comprehension and mental lexicon (see [Culter, 1983], [Davidson and Green, 1988], [Frazier, 1985], [Frazier, 1988], and [Klare, 1963].

Semantic Complexity

Currently, we have little to say about Semantic Complexity. However, it will deal with those difficult topics such as pragmatic, discourse analysis, metaphor, anaphora, ellipsis, semantic ambiguity, PP attachment, and so on.

4 The Paraphraser

The Paraphraser is the first module in the proposed model and can be considered as a preprocessor to a traditional MT system. The paraphraser is in charge of transforming Class II sentences into Class I. It applies paraphrasing rules to reduce the lexical, syntactic, and semantic complexities of a natural language sentence.

Let us suppose that in the sentence complexity formula, TranX, LexX, SynX, and SemX range between 0 and 10 and the boundaries between classes are:

$$0 \leq Class\ I < 3 \leq Class\ II \leq 6 < Class\ III \leq 9 < Class\ IV \leq 10$$

Then, a sentence is in Class II if $3 \leq TranX \leq 6$, i.e., at least (the largest) one of LexX, SynX, and SemX is between 3 and 6. And, the Paraphraser is supposed to reduce all LexX, SynX, and SemX to under 3. The reducing process is not necessarily monotonic. For example, a sentence(with LexX:2, SynX:5, SemX:4) can be paraphrased bt reducing the TranX from 6, 5, 4, and finally to 3. Note that LexX is increased temporarily from 2 to 3 after Paraphrase 1.

The complexity-reducing process can be viewed as a minimizing process in the complexity-space towards the origin (or a minimum). A paraphrasing rule can be considered as a transition vector toward the minimum.

Some Examples of Paraphrasing Rules

In this section, we give some examples of paraphrasing rules (P-Rules) to illustrate a possible implementation of the paraphraser. In the examples, the P-Rules are represented in the form of productions, rewrite rules, or situation/action rules.

With the help of the lexicon, a simple rewriting process can be used in paraphrasing.

Example 1

First, we take an example from Lehrberger [1986]:

	Refill	. tank	if	fluid level	low.
L':	V	N[sg,count]	SC	NP	A
L_{std}:	V	DET N[sg,count]	SC	NP	BE A

where V stands for verbs, N[sg,count] singular countable nouns, SC subordinate conjunctions, NP noun phrases, A adjective, DET deteminers, and BE be-verbs.

To paraphrase the given sentence from L' to L_{std}, we design two P-Rules as follows:

P-Rule 1: V N[sg,count] $\Longrightarrow$ V DET N[sg,count]

P-Rule 2: SC NP A $\Longrightarrow$ SC NP BE A

The BE in P-Rule 2 can be either "is" or "are" according to the number of the NP. As Lehrberger mentioned, the DET in P-Rule 1 can be either "the" or "a" depending on the context.

Example 2: Conjunctions in Chinese

The corresponding word of "and" in Chinese is he or yu. For the purpose of compactness, the conjunctions are usually omitted in the sublanguage. An entry in our corpus is Jiang ZeMin tan tai gong mei zhong wenti. which means "Jiang ZeMin talked about the problem among Taiwan, Hong Kong, America, and China.". However, no punctuation, no conjunction, and no preposition (actually even no blanks between words) exists to mark the relation of conjunction in the original sentence. (By the way, no tense marker or article is required in Chinese.) For the phrase "among Taiwan, Hong Kong, America, and China", all we have is a list of four (single-character) words tai gong mei zhong. (The four words are shorthands for taiwan xianggong meiguo zhongguo. But it is not important in this example.) Paraphrasing rules can be used to mark the relation. (We assume the P-rules are ordered so the rule in the front will be fired first.)

P-Rule 3: NP[name] NP[name] NP[name] NP[name] $\Longrightarrow$ NP[name], NP[name], NP[name], he NP[name]

P-Rule 4: NP[name] NP[name] NP[name] $\Longrightarrow$ NP[name], NP[name], he NP[name]

P-Rule 5: NP[name] NP[name] $\Longrightarrow$ NP[name] he NP[name]

5 Conclusion

In this paper, we have reviewed the concepts of sublanguage and linguistic complexity. Then, the PBT model for machine translation is presented. We also discuss the translatability and translation complexity and identify the significance and function of the Paraphraser. After that, some examples are given to illustrate the usage of paraphrasing rules.

References

[Becker, 1975] J.D. Becker. The phrasal lexicon. In R. Schank and B.L. Nash-Webber, editors. *TINLAP-1*, Cambridge, MA, 1975.

[Bloom, 1981] A.H. Bloom. *The Linguistic Shaping of Thought.* Lawrence Erlbaum, Hillsdale, NJ, 1981.

[Culter, 1983] Anne Culter. Lexical complexity and sentence processing. In G.B. Flores d'Arcais and R.J. Jarvella, editors, *The Process of Language Understanding*, chapter 2, pages 43–79. John Wiley and Sons Ltd., 1983.

[Davidson and Green, 1988] A. Davidson and G.M. Green, editors. *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered.* Lawrence Erlbaum, Hillsdale, NJ, 1988.

[Frazier, 1985] Lyn Frazier. Syntactic complexity. In D.R. Dowty, L. Karttunen, and A.M. Zwicky, editors, *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, chapter 4, pages 129–189. Cambridge University Press, 1985.

[Frazier, 1988] Lyn Frazier. The study of linguistic complexity. In A. Davidson and G.M. Green, editors, *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, chapter 8, pages 193–221. Lawrence Erlbaum, Hillsdale, NJ, 1988.

[Grishman and Kittredge, 1986] R. Grishman and R. Kittredge, editors. *Analyzing Language in Ristricted Domains: Sublanguage Description and Processing.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[Kittredge and Lehrberger, 1982] R. Kittredge and J. Lehrberger, editors. *Sublanguage: Studies of language in restricted domains.* Walter de Gruyter, Berlin, 1982.

[Klare, 1963] G.R. Klare. *The Measurement of Readability.* Iowa State University Press, Ames, Iowa, 1963.

[Lehrberger, 1986] J. Lehrberger. Sublanguage analysis. In R. Grishman and R. Kittredge, editors, *Analyzing Language in Ristricted Domains: Sublanguage Description and Processing*, pages 19–38. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[Liang, 1987] S.-C. Liang. *Far East Concise Chinese-English Dictionary, 2nd Ed.* The Far East Book Co., Ltd., Taipei, 1987.

[Nirenberg, 1987] S. Nirenberg, editor. *Machine Translation: Theoretical and Methodological Issues.* Cambridge University Press, Cambridge, 1987.

[Sager, 1972] N. Sager. Synatctic formatting of science information. In *AFIPS Conference Proceedings, 41*, pages 791–800, 1972. Reprinted in [Kittredge and Lehrberger, 1982].

[Slocum, 1987] J. Slocum, editor. *Machine Translation Systems.* Cambridge University Press, Cambridge, 1987.

Kenneth W. Church

USING STATISTICS IN LEXICAL ANALYSIS

The computational tools available for studying machine-readable corpora are at present still rather primitive. In the Cobuild project of the 1980s, for example, the typical procedure was that a lexicographer was given the concordances for a word or group of words, marked up the printout with colored pens in order to identify the salient senses, and then wrote syntactic descriptions and definitions. Although this technology is an advance on using human readers to collect boxes of citation index cards (the method Murray used in constructing the Oxford English Dictionary a century ago, and still in use in some present-day lexicographic organizations), it works well only if there are no more than a few dozen concordance lines for a word, and just two or three main sense divisions. In analyzing complex words such as \fIstrong\fP and \fIthat\fP, the lexicographer is trying to pick out significant patterns and subtle distinctions that are buried in literally thousands of concordance lines. The unaided human mind simply cannot discover all the significant patterns, let alone group them and rank them in order of importance. Concordance analysis is still extremely labor-intensive, and prone to errors of omission.

Computational linguists run into similar problems when they try to write grammars, especially disambiguation rules to diagnose lexical ambiguity. Again, the most common words cause the most trouble. Consider the word \fIthat\fP, which has many parts of speech, and is frequently found in expressions such as: \fIthat way, that moment, that point, so that, think that, now that, indicated that, all that\fP, etc. Like concordance analysis, designing disambiguation rules is still extremely labor-intensive, and prone to errors of omission. The unaided grammar writer simply cannot discover all the significant patterns, let alone group them and rank them in order of importance.

This talk will discuss a number of statistical tools and show some examples of how these tools can be used to enhance productivity of a lexicographer or a computational linguist. It will be assumed that we can depend on human judgment to use the statistics appropriately, and to check the results to make sure that they are reasonable. The emphasis on human interaction distinguishes our approach from self-organizing approaches.

E. Dermatas, G. Kokkinakis

AN ALGORITHM FOR AUTOMATIC GRAMMATICAL CLASSES DEFINITION

1. Introduction

Systems based on Markovian processes for the prediction of grammatical classes in written text have a good performance and an acceptable response time in low memory requirements, but are still far from a good performance, real-time response and acceptable costs in high number of grammatical classes requirements.

In this paper an algorithm for automatic grammatical classes definition from a detailed system of grammatical categories is presented. The algorithm recursively optimizes a double criterion.

(a) The prediction quality of the Markov model.
(b) The information obtained from the predicted grammatical classes.

The performance of the algorithm has been measured on a newspaper text of 120000 words. These words were initially manually labelled using a very large set of grammatical labels (approximately 1200). From this set three sets of 90, 200 and 500 grammatical classes were automatically extracted. Then each of the above sets was reduced to three sets of 5, 15 and 20 grammatical classes using the proposed algorithm.

The paper has the following structure:

A theoretical approach to the problem of defining an optimum set of grammatical classes is presented first. A detailed description of the algorithm which defines the sets of labels for the selected criterion follows. The environment, the description of the experiments and the measurements with comments for the performance of the algorithm are described in the last part of the paper.

2. Theoretical Approach

Let $w_1, w_2, ..., w_T$ be the words of a natural language text which have been labelled as $g_1, g_2, ..., g_T$ using a set of labels:

$$G = \{ G_1, G_2, \dots , G_k \} \tag{1}$$

If we define a function:

$$tg: G \rightarrow G' \tag{2}$$

where

$$G' = \{ G'_1, G'_2, \dots , G'_L \} \tag{3}$$

with

$$L < N \tag{4}$$

the transformed sequence of labels in the text is taken from the relation:

$$W' = \{ w'_i = tg(w_i) , \ i = 1,N \} \tag{5}$$

The function tg(.) is defined using two criteria in the new set of labels W':

(a) The Markovian model must be efficient in the prediction of the classes of G' in the text.
(b) The information given by the states (classes) of the model must be maximized.

The corresponding quantities of the above criteria are chosen as follows:

(a) The efficieny of the Markov model is measured by the negative natural logarithm of the perplexity. The perplexity is defined as the probability of the correct sequence of labels in the text and is given by the relation:

$$L'p = -\frac{1}{T} \log_2 P(g'_1, g'_2, \dots , g'_T) \tag{6}$$

With the hypothesis of the probabilistic independence of labels which have position distances more than k, the relation (6) is transformed to:

$$L'_p = -\frac{1}{T} \log_2 \left(\prod_{i=1}^{t-k} P(g'_i \mid g'_{i+1}, \dots , g'_{i+k}) \right) P(g'_{T-k+1}, \dots , g'_T) \tag{7}$$

If the training text has a great number of word entries (T >> 1) the above formula can be rewritten as:

$$L'_p = -\frac{1}{T} \log_2 \prod_{i=g'_a,g'_{b_1},\dots,g'_{b_k}} P(g'_a \mid g'_{b_1}, g'_{b_2}, \dots , g'_{b_k})^{f(g'_a \mid g'_{b_1},\dots,g'_{b_k})} \tag{8}$$

where:

f(.) is the frequency of occurence of the sequence of grammatical labels.

P(.) is the probability of occurence of ga given the sequence of $g_{b1}, g_{b2}, ..., g_{bk}$ which is estimated from the training text.

The multiplication factor contains all the sequences which occured in the training text.

Finally, the relation of the negative natural logarithm of the perplexity is given by the formula:

$$Lp(G') \simeq -\sum_{a,b_1,\dots,b_k} \log_2 P(g'_a \mid g'_{b_1} , \dots , g'_{b_k}) \tag{9}$$

When Lp(G') decreases for the specific definition of G', the prediction quality of the Markovian model is increased.

(b) The quality of the information given by the classes G' is taken from by the formula of entropy:

$$E(G') \simeq -\sum_{i=1}^{L} P(G'_i) \log_2 P(G'_i) \qquad (10)$$

For a specific set of G' the quantity of information given by the prediction model increases when the corresponding entropy of G' is increased.

The total criterion for the optimum set of G' is chosen to be the set G_0' defined by the relation

$$G'_0 = \underset{G'}{argmin} \frac{Lp(G')}{E(G')} \qquad (11)$$

An analytical solution for the globally optimum set G_0' has not yet been estimated. Nevertheless, a local optimum can be obtained.

3. Description of the Algorithm

The following algorithm estimates a local optimum set.

Step 1. Initial definition of the classes G'(0), I = 0.

Step 2. I = I + 1.

Step 3. For the Gj label we compute the quality measures when the label is moved to class 1, for all classes.

$$P_{jl}(l) = \frac{P_l(G_j \varepsilon G'_l(l-1))}{E(G_j \varepsilon G'_l(l-1))} \qquad (12)$$

Step 4. Definition of the label and the set which gives the minimum criterion.

The set G'(I) is defined from the G'(I-1) when the label G_{j0} is moved into G'_{l0}.

$$(l_0, j_0) = \underset{l,j}{argmin}\ P_{lj}(l) \qquad (13)$$

$$P(l) = \min_{i,j} P_{ij}(l) \qquad (14)$$

If P(I-1) - P(I) > Mi and I < I_{max}, the algorithm is repeated from step 2.

Step 5: the set G'(I) is a local optimum solution.

4. Environment - Experiments

The performance of the above algorithm has been measured in a series of experiments. A set of 120000 words from a newspaper text has been semi-automatically labelled using a set of approximately 1200 grammatical labels.

Three sets of 90, 200, 500 initial labels extracted from the above set and three sets of 5, 15, 20 final classes have been used to test the behaviour of the algorithm.

The measured performances an excerpt of which is shown in the folllowing table consist of the following quantities:

(a) Response time of the algorithm.
(b) Perplexity of the Markovian model.
(c) Entropy of the predicted grammatical labels.
(d) Number of transitions.

The software has been written in C language and the experimental results were obtained by a VAX 3100 computer.

The basic results from the experiments can be summarized as follows:

(a) The algorithm converged into a local optimum in all experiments.
(b) A good local optimum was reached in the majority of the experiments.
(c) Generally the algorithm in the initial steps classified the most probable labels and finally transferred the less probable labels into the correct classes.
(d) The number of the algorithm steps increased when the number of initial or the number of final labels was increased.

A number of experiments investigated the relation between the initial definition of the classes and the local optimum estimated by the algorithm.

More detailed results will be presented in the paper.

5. Conclusion

The algorithm presented in this paper is an efficient tool for a well defined set of grammatical labels it ensures:

(a) A small number of transition probabilities.
(b) Good prediction accuracy of a Markov model.
(c) High quality of the information given by the model.

Further improvements of the algorithm can be obtained:

(a) Using a different criterion. The above criterion has been chosen from three others as giving the best performance. A set of different criteria will be tested in the near future.
(b) Changes in the algorithm structure can decrease the computing time and/or lead to convergence to a global optimum.
(c) A study of the relation between the initial definition of the classes and the final local optimum can be reached by a criterion which defines an initial set which converges to a "good" local optimum or to the global optimum set.

INITIAL	LABELS	PERPLEX.	ENTROPY	RATIO	TRANSIT
1	185/10	2.1724	1.9422	1.1837	182
2	185/15	2.2451	1.2473	1.1928	225
3	185/20	2.9408	2.4286	1.2102	381
FINAL	LABELS	PERPLEX.	ENTROPY	RATIO	TRANSIT
1	10	2.2597	2.1847	1.0341	99
2	15	2.0598	2.4893	0.8272	113
3	20	2.0586	2.6227	0.7843	138

Methodological Aspects of the Categorization of a Middle French Corpus with a Computer-Assisted Text Analysis Software (SATO)

Fernande Dupuis
Daniel Gosselin
Centre d'analyse en syntaxe historique
Département de linguistique
Université du Québec à Montréal
C.P. 8888, Succ. A
Montréal (Canada)
bitnet: R15760@UQAM

Benoît Habert
École Normale Supérieure de Fontenay Saint Cloud
31 avenue Lombart
F-92260 Fontenay-aux-Roses (France)
internet: bh@litp.ibp.fr

1. Introduction

Last decade's achievements in computer science have had major implications for diachronic linguistics as well as for other linguistic disciplines. In our domain, historical syntax, we cannot perform experiments to broaden the range of available data. Since the study of syntactic change in the time course of a given language is highly dependent upon the availability, for analytical purposes, of extended corpuses, the use of computational tools has by now become a necessity.

In the following discussion we will focus on the contribution to the statistical study of an archaic language, Middle French (henceforth MF), of a linear, word-by-word categorization tool (without phrase construction nor strategies for dealing with ambiguities).

2. Construction of a corpus for the quantitative study of syntactic phenomena in an archaic language

A crucial problem in diachronic syntax is that data are confined to those which survived the vicissitudes of time. To compensate for sampling biases, we cannot obtain information by eliciting acceptability judgments from informants, as it is routinely done when studying a living language (Krock 1990). Furthermore, getting negative evidence is a result that even an extremely acute knowledge of the grammar would hardly allow one to achieve.

Another problem is raised by the fact that although most existing grammars provide quite satisfactory syntactic description, they do not always distinguish unusual structures from more familiar ones (for the case of MF, see, e.g., Marchello-Nizia 1979, Zwanenberg 1978, Nissen 1943, Price 1973, Martin 1978, Martin & Wilmet 1980, Dees 1978). Therefore, reliance on statistical evidence becomes critical in determining the salient linguistic tendencies and their average effect on certain areas of syntax, as well as the status of rare phenomena.

The problems reported so far emphasized the limits of a research based on data sampling and pointed to the need for extensive corpuses. Furthermore, the construction of a corpus for the study of historical materials must be carefully planned. For instance, considering the syntactic properties described below, a suitable corpus of MF should include data from both prose and verse, from different dialectal area and from different time periods.

3. SATO

To recollect the relevant linguistic data of the corpus we turned to SATO, a computer-assisted text analysis software written in C-language by François Daoust of the *Centre d'analyse de textes par ordinateur* of *UQAM*, which runs on IBM compatible microcomputers. Its main purpose is to permit a lexical and statistical analysis of texts.

Considering the immediate goals of our research, one of the main reasons which motivated the choice of SATO was the decision not to develop from scratch a new computational tool. To this we must add our limited knowledge of alternative tools; as far as we can judge, such a limited knowledge is characteristic of the dialog between linguistics and the computational sciences (further problems include the absence of a real organized market as well as the absence of standards). Moreover, for practical reasons, we were looking for an easy-to-use software program running on a PC.

SATO allows one to assign a given number of abstract properties to lexemes or parts of the text. It works in two steps. First, one must provide the text in an ASCII form to the sub-program SATOGEN, which creates secondary files for the analysis of the text without modification of its content.

The second step requires the use of SATOINT, another sub-program of SATO, which allows the user to ask various questions regarding the text. For example, it can assess the readability of the text, evaluate the number of occurrences of a given form, and provide concordances of different forms. Twp types of displays are available: on the screen and in an output file.

In our case, SATO was mainly used to assign a chosen lexical element a number of grammatical or morphological properties relevant to the analysis of the sentences of a text. To do so, symbolic properties were attributed to that element of the clause.

Once these properties had been distributed, a statistical account was made of the relative proportion of each property. This provided an overview of the linguistic particularities of a given text. SATO was subsequently used to list specific properties, alone or in relation with other properties or words in the text.

Furthermore, SATO permitted the simultaneous processing of several texts, thus providing overviews of the whole corpus or of subunits of the corpus. In addition, variation between texts, because of the author or of the period in which they were written, could be easily spotted.

SATO, while certainly not a "perfect" tool, has until now met a good part of our vital analytic needs. Among its drawbacks, we can point to the impossibility of executing changes in the original text. Also, since SATO is not a small parser, it cannot assign properties to given contexts. Property assignment must be done beforehand, with the help of a word processor.

SATO relates graphic forms, devoid of linguistic reality, to data organized on a linguistic basis; it does so by attaching arbitrary long sets of properties/values to lexemes, either linearly on the text (as we did in our research) or by projecting a lexicon whose elements are tagged with a category index, or finally by completing the linear process with the result of the projection of a previously indexed lexicon. Finally, SATO can also introduce new segmentations.

4. The Subject in MF

In a nutshell, among the most striking properties of MF, a stage of French which roughly extends from XIVth to XVIth century, is the fact that the subject pronoun is often missing in declarative clauses. Despite this apparent similarity with Romance languages, such as Italian or Spanish, one fundamental difference sets MF apart from these languages: the missing subject construction is highly selective and appears only when the inflected verb is preceded by any type of constituent, including a null-operator. Such a state of affairs is reminiscent of the familiar Verb-second constraint found in most Germanic languages.

The absence of the subject pronoun has often been linked by the grammatical tradition to a verbal morphology rich enough to recover the semantic content of the subject. Since verb endings tend to neutralize in MF, the subject pronoun should be expected to obligatorily appear. Unfortunately, this assumption does not match our present data.

We have chosen a linear categorization of various significant properties/values on each finite verb of the text, such as the presence or absence of the subject pronoun related to the person features of the subject. This has led us to a quantitative analysis according to which the phenomena underlying the absence or presence of the subject has little to do with rich inflection on the verb.

The results from our corpus (which comprises more than 8000 finite verbs) show, for instance, that the 3rd person (singular and plural) is omitted at the rate of 33% and 34% respectively, while with the more marked 2nd person plural the percentage of omitted subject is 37%. From these facts, we see no compelling evidence to sustain the above assumption.

However the proportion of omitted subject raises to 50% when it is the impersonal 3rd person singular. The statistical facts suggest that the behavior of the subject is more closely linked to its thematic role in the sentence rather than to the functional features of verbal endings.

5. Pushing the limits

Among the not yet fully exploited potentialities of SATO, the following should hopefully be implemented in a near future:

- the possibility to abstract part of the linear categorization already executed into a lexicon which can be mapped onto other texts speeding up the descriptive category recognition process; we are presently working on this aspect.

- the immediate quantification of "linear" phenomena onto combinations of properties, reordering them until relevant facts have been found. All of this has to do with the possibility of comparing different parts of the corpus.

- the verification of predictions: knowing that certain combinations are not possible, SATO should search for those particular combinations.

As a more general point, it would be fruitful to build upon the heuristic aspect of the construction of properties; this implies the creation of an analytical grid shared by the team of researchers.

6. What a simple categorization tool must integrate

In spite of an obvious need for improvements in SATO's interface (such as: clear menus, mouse, friendly input/output), we are still uncertain at the moment whether SATO should be abandoned and replaced by a more powerful tool or not. The best option seems to consist in pushing to the limits what a linguist can expect (and obtain) from a categorization tool similar to SATO.

For linguists having to choose between different softwares we could mention these additional criteria.

SATO provides only structureless informations on properties: the retained form does not induce any formalization on the lexicon, relation between properties (logical dependencies, redundancy, etc.). Notice that more formalized systems (cf DAGS from unification formalism) show similar limits. (But see contra the propositions on hierarchy from HPSG [Pollard & Sag 87, p 191-218]). Following propositions from [Gazdar et al. 88] it appears to be necessary to define syntax and semantics for a system of given categories.

SATO offers a limited number of properties. The problem is getting more acute as the results of the analysis create the need for new properties. From a general point of view the system of categories must evolve in a coherent way from one stable stage to another.

Verification tools are needed:
- for the coherence of the syntactic or semantic list of properties
- for the coherence of the the tagging work of one researcher
- for coherence across researchers.

As a general remark the utilization of SATO raises the question of how to build upon the expertise developed during the categorization process: feed-back on the categories, missing features, etc. The application should include a notebook-like tool on which fragments of text raising specific problems could be preserved. In particular the notebook could be used to store the tests formulated during the categorization process. For the moment we have only included a property allowing the marking out of questionable decisions during categorization.

Finally, a new user should not be immediately confronted with the complexities of the entire system of categories; rather s/he should be able to proceed from the most evident properties, associated with easy linguistic tests, to the most difficult properties. Moreover, empty properties should be provided for the most advanced users.

PROPOSAL 'CONSTRAINT MEASURE' FOR VARIOUS KINDS OF TEXT

Koichi Ejiri
Adolph E. Smith
California Research Center
Ricoh Corporation
2882 Sand Hill Road, Suite 115
Menlo Park, CA 94025

ABSTRACT

Zipf's law is one of the most famous generalizations in text statistics. S. Mizutani introduced a new representation which is more appropriate for detailed analysis of text. We prove that his representation is well approximated by the ratio between a number_of_words and a number_of_different_word in the target text. This ratio was applied to several examples of text; computer language, non-native English, and standard English. We found that his representation, which is approximated by the above ratio, is modified into a better measure of constraint of the sentence; G = log(N/L)/{log(N) - 1}, where N is the number of words in the text, L, the number of different words. This measure roughly classifies the English text in the order of constraint. Here, the constraint means the size of the domain expressed by the text.

We also found that this 'G' score in English reading text books for elementary school in the United States decreases during first and second grade. Then reaches to the bottom about seventh grade. The similar tendency is observed for Japanese texts also.

INTRODUCTION

An intensive effort has been taken to develop a technology to understand a natural language (NL) text all over the world. Some systems claim that they can understand natural language. However, the applicability of the NL technology is heavily dependent on the nature of the given text. There is no single system which can understand general English texts. To focus on narrower target, there is a strong demand to categorize the given text for further processing.

Interest in identifying or classifying text has a long history, especially for cryptographic applications. One of the simplest approaches is the use of the n-gram frequency which has a strong tendency to reflect author's preference. Some software has been available for this purpose(Ref.1).

On the other hand, there has been some effort to generalize rules from the texts. In 1935, Zipf introduced a generalized formula which can express the distribution of word frequency regardless of language (Ref. 2). According to him the frequency of a word, f, and rank order of the word, r, were expressed as;

$$f * r = \text{constant}.$$

Here, rank order means that a word is ordered by its frequency, i.e. the highest frequency word comes first then the second largest frequency word is followed. This order is called rank order of the word.

Mizutani (Ref.3) mentioned that this formula had been discovered by both J.B.Estoup(1916) and E.V. Condon (1926), independently. In his book, he proposed a new formula;

$$K = L * f / (a * f + N * b) \qquad (1)$$

where the symbols are

K; total number of different words whose frequencies are less than or equal to f,
L; total number of different words,
f; frequency of a word,
N; total number of words,
a, b; constants.

This can be converted to

$$y = f/K = \frac{a*f + N*b}{L} = c * f + d. \qquad (2)$$

Here, 'y' is a defined parameter, and both 'c' and 'd' are constants. It is easily understood that expression (2) represents the linear relation between 'f' and 'f/K'. Mizutani reported that the relation (2) is valid for various languages, including Japanese, English and European languages (Ref.3).

EXPERIMENTS

Mizutani also mentioned (Ref.4) that in application of relation (2), we should consider both morphological change and the part of speech; for example, 'work(verb)' and 'works(verb)' should be treated as the same word, but 'work(noun)' and 'work(verb)' should be treated as different words. However, to do so, it requires a lot of processing including a large dictionary search. If the expression (2) is stable against the morphological changes, this could be a more practical method.

To test this validity, we compared two statistics for various English texts.

In this experiment, a word is defined as; "the character string between the dividers" where dividers are defined below:

<Dividers>

(space)	'	/
(	)	!
@	^	=
;	[	]
*	%	+
{	}	&
#	|	<
>	-	:
"	.	?
(EOL)	(CR)	\

Here, .(EOL) and (CR) are "end of line" and "carriage return", respectively.

The comparison was made between Mizutani's expression (equation 2) and 'Zipf's second law'.

$$k * (f * f) = \text{constant} \qquad (3)$$

where, k is the number of the words which have the same frequency value 'f'. A more generalized form was given by Mandelbrot (Ref.6).

For simplicity, every different spelling is treated as a different word; i.e. plural form of nouns, verb inflection or mis-spelled words are identified as different words. The tested results are shown in Fig. 1 and 2. According to expression (3), Fig.1 is expected to be a flat curve, but is not. Expression (2) indicates that y is proportional to f. From our results, it is obvious that Mizutani's expression is more accurate in comparing the different samples. Because, camparison of single parameters, i.e. slope of the curve in Fig.2 in this case, is easily done for various texts. For this reason, only Mizutani's expression will be discussed hereafter.

Data in both Fig.1 and Fig.2 are obtained from the following different texts;

(c): computer language, C or Fortran source code (graphic symbol(G.S.) = +)
(t): trouble shooting description for a facsimile; trouble shooting knowlwge (G.S.= ×)
(f): English text written by foreigner (G.S.= ◇)
(n): English written by a native (G.S.= ✱)
(m): Manual for computer software (G.S.= □)

In Fig.2, 'frequency' was normalized for 1,000 sampled words. The line in this figure is determined from the linear regression approximation (2).

Another finding from this graph is that the slope 'c' in expression (2) of those lines are strongly correlated to the constraint of expression. Because the tested texts are roughly placed in the constraint order, (c) > (t) > (m) > (f) > (n). Although, constraint does not have a clear definition, it is reasonable that the diversity of the world to be expressed by the texts are inversely proportional to the constraint. It is clear that the target world becomes larger in the order of (c) -> (t) -> (m) -> (f) -> (n).

REFERENCES

(1) Tankard, J; "The literary detective", Byte 1986, February, pp231

(2) Zipf, G.K. "The psycho-biology of language", The MIT Press (1965), Originally printed by Houghton Mifflin Co. 1935
Nicolis, J.; "Dynamics of Hierarchical Systems", Springer-Verlag, Berlin 1956

(3) Mizutani, S; "Lecture on Japanese", Asakura, Tokyo

(4) Mizutani, S; Private communication, Apr. 1989

(5) Howes, D. "On the relations between the Probability of Word as an Association and in General Linguistic Usage," J. of Abnormal and Social Psychology, vol.54, pp.75-85, Jan. 1957

(6) Mandelbrot, B.B. "Fractal Geometry of Nature", W.H.Freeman and Co., New York, 1982

(7) Hoermann, H. "Psycholinguistics", translated by H.H.Stern and P.Leppmann, Springer-Verlag, New York, 1979

(8) Booth, A. D. (1967) A law of Occurrences for Words of Low Frequency, Information and Control, 10(4); 386-393

Sheila Embleton

MULTIDIMENSIONAL SCALING AS A DIALECTOMETRICAL TECHNIQUE: OUTLINE OF A RESEARCH PROJECT

Dialectometry is the study of quantitative measures of distance between dialects. It is the "application of the principles of numerical taxonomy to the analysis of dialect data" (Schneider 1984, 314), and "aims at the recognition of patterns by means of numerical classification" (Goebl 1984, iii). As such, it goes beyond, for example, the simple use of a computer to draw dialect maps, to the real use of mathematical/statistical methods in classification and analysis. Dialectometry has the usual advantages of mathematical techniques (e.g., "objectivity", speed, ability to handle large volumes of data) as well as the usual disadvantages of mathematical techniques (e.g., loss of detail when dealing with aggregated data); it is intended to supplement/complement more traditional methods in dialectology, not to replace them.

Any dialectometrical technique must deal with three issues: how to construct the distance measure (D), what to do with the resulting numbers, and how to interpret the results. The first of these is relatively independent of the technique used and there are fairly standard measures in use; hence this issue will not be discussed here. This neglect should not be taken to mean that there are no problems involved here, and one of the aims of this research project is to investigate some of these problems and suggest some solutions. The second and third issures are intimately bound to the dialectometrical technique used, and hence must be discussed together with that technique.

There have been at least seven dialectometrical techniques discussed in the literature. Some of these have been more successful than others, but it is not the aim of the current paper to discuss or evaluate these. The current paper presents á further technique, one involving multidimensional scaling (MDS), and shows that it has some advantages and some disadvantages compared to other techniques. It schould be viewed as adding another dialectometrical tool for researchers, not as replacing existing tools.

MDS refers to a collection of statistical techniques for representing the similarities among a set of objects spatially, as for example 2-dimensionally in a map. Like factor analysis, MDS provides a means of compressing a large mass of data by representing similarity in terms of a small number of underlying dimensions. Unlike factor analysis, MDS makes only weak assumptions about the data; thus MDS is more appropriate than factor analysis for linguistic data (as well as many other types of data in the social and humanistic sciences). MDS proceeds by taking data originally in k dimensions, and reducing it successively to k-1 dimensions, and so on, down to 2 dimensions or even just one dimension. There are commonly available statistical package programs for MDS.

We first construct a matrix of values of D, pairwise for each pair of dialects being compared; this matrix will be in as many dimensions as we have varieties/locales under investigation. Using MDS, we can reduce the number of dimensions to 2, which gives us a "linguistic map" of our dialect space. This can then be compared to the locations of the varieties on a regular "geographical map", which is after all just a conventional 2-dimensional representation of geographical distance. In our comparison of the linguistic map with the geographical map, we are looking for discrepancies (i.e., differences) between the two maps. It is in the interpretation of

these discrepancies that our interest actually lies. MDS has been previously applied in the case of different languages by Black (1976; to 3 language families) and by Dobson and Black (1979; to Australian languages) with enough success to warrant its investigation as a tool in dialect study.

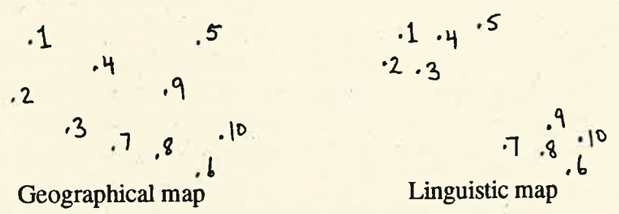
Compared to true geographical distance, we would expect linguistic distance to be distorted in two possible ways. Linguistic distance between two varieties can be increased (compared to geographical distance) by the existence of isogloss bundles between the two locales, or it can be decreased by intercommunication between the two locales. Thus a dialect boundary will have the effect of pushing the two locales further apart on the linguistic map, whereas intercommunication will have the effect of pulling the two locales closer together. The first of these two situations is illustrated (in a hypothetical case) by Map 1, and the second (also in hypothetical case) by Map 2.

Map 1

Map 2

To take a larger, but still hypothetical, case, consider a situation such as the following, illustrated in Map 3.

Map 3



The discrepancies between the geographical map and the linguistic map would suggest that there is an isogloss bundle between locales 1,2,3,4, and 5 on the one hand, and locales 6,7,8,9, and 10 on the other hand. We could probably also reasonably conclude that locales 1 and 6 are important centres (e.g., centres of trade, culture, or politics), because they appear to act like magnets in pulling other locales on the linguistic map closer in to their "field" of influence.

Compared to other dialectometrical methods (e.g., those of Séguy and Goebl), this dialectometrical method has some advantages and some disadvantages. The advantages are that the maps are easily and quickly constructed by computer and that all the information is integrated onto the one map in an immediate and obvious way. Isogloss bundles and "rapprochements" are shown clearly, and in a visually appealing way. There is one disadvantage compared to Goebl's method (involving choroplethic maps), which is not so much (in my view) a disadvantage as a "trade-off". Goebl uses one map per locale, which, although it doesn't integrate the data as well, preserves more detailed information about each locale and its relationship to all other locales. The MDS method integrates the data more, but in return for this one loses much of the detail of the picture. This, however, is the type of trade-off often faced by the user of statistical methods in the social and humanistic sciences - how much does one want to see the forest vs. how much does one want to see the trees vs. how much does one want to see the individual leaves on the tree? It is not so much a question of one method being "better" or "more useful" than another, as of different methods illuminating different aspects of the same problem, or perhaps being more appropriate to different kinds of problem. Another advantage to MDS, not yet investigated by the present author, but implicit in some of the dual-scaling work of Cichocki and Flikeid, is that extra dimensions can easily be added to the data. Thus, if the data were available for each locale, a sociolinguistic dimension could be added, and perhaps the linguistic representation resulting from MDS could then be a 3-dimensional spatial configuration - literally adding a sociolinguistic dimension to the geographical dimension.

The MDS method has been tested (in a pilot-study) on data (128 lexical items) from 18 representative locales (out of 313) in Orton and Wright 1974. The results (maps to be presented at the conference) show a good match between the geograhical map and the linguistic map, but, as explained above, it is from the discrepancies between these two maps that our information about the dialect situation comes. The linguistic map shows, among other things, tighter clustering of the locales into a Northern, a Midland, and a Southern group, as well as a noticeable effect of London as a "magnet". These coresspond to results from "traditional" dialectology. In addition to values of D based on lexical data, phonological (203 items), morphological (76 items), syntactic (9 items), and "overall" (based on all the data plus a further 59 lexical items) distances were also computed for the same 18 locales, using Orton et al. 1978, and subjected to the MDS technique. The phonological, morphological, and "overall" data all produced essentially the same distortions (and hence the same conclusions) as the lexical data discussed above. The syntactic data, with only 9 items, was insufficient to justify the MDS technique (and the resultant map was essentially random and meaningless, as expected).

Using MDS may become more difficult as the number of locales investigated increases, due to the limitations of the human eye. The pilot-study may have worked well simply because it involved only 18 locales. With larger data sets, it may work only where the patterns to be extracted from the data are so strong that, no matter what one does, the answer will be glaringly obvious, and one wouldn't have needed sophisticated statistical techniques anyway. Maybe one will have to select fewer "representative" points (with all the implications that that has for "objectivity"), or only use the method for certain types of dialect situation. In order to investigate these issues, a research project involving the testing of this MDS dialectometrical technique on all of the data from all 313 locales has recently been approved for funding by the Social Sciences and Humanities Research Council of Canada.

REFERENCES

Babitch, Rose Mary (Ed.) (1988): Papers from the Eleventh Annual Meeting of the Atlantic Provinces Linguistic Association. Centre Universitaire de Shippigan, Université de Moncton: Shippigan, New Brunswick.

Black, Paul (1976): "Multidimensional scaling applied to linguistic relationships", in: Cahiers de l'Institut de linguistique de Louvain 3, 43-92.

Cichocki, Wladyslaw (1989): "An Application of Dual Scaling in Dialectometry", pages 90-96 in: Kretschmar et al. (1989).

Dobson, Annette J., Paul Black (1979): "Multidimensional scaling of some lexicostatistical data", in: Mathematical Scientist 4, 55-61.

Flikeid, Karin, Wladyslaw Cichocki (1988): "Application of Dialectometry to Nova Scotia Acadian French Dialects: Phonological Distance", pages 59-74 in: Babitch (1988).

Goebl, Hans (Ed.) (1984): Dialectology. Bochum (Brockmeyer).

Kretzschmar, William A., Jr., Edgar W. Schneider, Ellen Johnson (1989): Computer Methods in Dialectology. Special issue of Journal of English Linguistics 22.1. (April 1989).

Orton, Harold, Nathalia Wright (1974): A Word Geography of England. London (Seminar).

Schneider, Edgar (1984): "Methodologische Probleme der Dialektometrie", pages 314-335 in: Goebl (1984).

Séguy, Jean (1971): "La dialectométrie dans l'atlas linguistique de la Gascogne", in: Revue de linguistique romane 37, 1-24.

Statistical Language Modelling using Cache Memory

Ute Essen, Hermann Ney

Philips GmbH Forschungslabor Aachen
P.O.Box 1980, D-5100 Aachen, Germany

Typical applications for automatic speech recognition systems, such as dictation systems, make great demands on the vocabulary size so that the language-model component becomes more and more important.

The task of a language model is to assign a probability to every possible sequence of words for a given vocabulary. The need for these prior probabilities results from Bayes' decision rule for minimum error rate. The word sequence $w_1 \dots w_M$ to be recognized from the acoustic sequence $x_1 \dots x_T$ is determined by

$$\max_{w_1 \dots w_M} \{ p(w_1 \dots w_M)\, p(x_1 \dots x_T | w_1 \dots w_M) \},$$

with $p(w_1 \dots w_M)$ being the probability that the word string $w_1 \dots w_M$ will be spoken (language-model part) and $p(x_1 \dots x_T | w_1 \dots w_M)$ the probability that when the speaker says $w_1 \dots w_M$ the acoustic evidence $x_1 \dots x_T$ will be observed (acoustic part). The probabilities of both parts guide the search of the recognizer in determining the final word sequence.

Using elementary rules of probability theory, the language model's probability can be decomposed as

$$p(w_1 \dots w_M) = \prod_{i=1}^{M} p(w_i | w_1 \dots w_{i-1}),$$

where $p(w_i | w_1 \dots w_{i-1})$ is the probability that w_i will be spoken, given that $w_1 \dots w_{i-1}$ were said previously. With a vocabulary of size W the computation of these conditional probabilities requires the estimation of about W^M different values. In practice, however, the training data is limited so that most histories $w_1 \dots w_{i-1}$ are unique or do not occur at all. In order to cope with estimation problems and storage requirements we restrict ourselves to a bigram language model.

We apply two methods of handling sparse data. The first one is a smoothing technique, which uses a weighted average of bigram and unigram events and leads to the interpolation formula

$$p(w|v) = (1-\alpha)\, q(w|v) + \alpha\, q(w),$$

where $q(w|v)$ is the probability of word w following word v, $q(w)$ the unigram probability of word w, and α the interpolation parameter.

The second one is the reduction of parameters by forming word equivalence classes (categories) using linguistic parts of speech or automatic clustering techniques. Assuming that the words w and v belong to one category only, we obtain the following formula with $q_1(c_w|c_v)$ being the conditional category probability of category c_w given category c_v, and $q_0(w|c_w)$ the probability of word w within its category c_w:

$$p(w|v) = q_0(w|c_w)\, q_1(c_w|c_v).$$

The bigram model is static because it only reflects the frequency distribution of words in the training set. In order to enable the adaptation to changing topics in testing data, we added a dynamic component, the so-called cache. The cache works as a kind of memory which includes the recent past of the word to be predicted into the estimation of its probability. We use the cache model in two variants:

1. The cache is part of the category model, as suggested and applied by Kuhn and de Mori [Kuhn 1990].

2. The cache is used for smoothing unigram probabilities in the word model.

The two variants were tested and compared.

In the category model, each class is provided with a cache of fixed length n which contains the n last words of the test set belonging to its category. The word probabilities estimated in training are modified by interpolation with the relative cache frequencies:

$$q_0^c(w|c_w) = \beta\, q_0(w|c_w) + (1-\beta)\, \frac{N^c_{c_w}(w)}{N^c_{c_w}},$$

where $N^c_{c_w}(w)$ is the frequency of word w in the cache of category c_w, $N^c_{c_w}$ the total number of elements in the cache of category c_w, and β the interpolation parameter.

The influence of the cache part on the quality of the language model was tested experimentally on two different corpora, a German corpus of newspaper articles and a more heterogeneous English corpus consisting of several different

text categories. As a measure of quality we chose the 'perplexity' which is the reciprocal of the average probability per word of sample text. Thus the better the model, the higher the probability it will assign to the sequence of words that actually occurs and, therefore, the lower the perplexity. For a more detailed examination of the cache effects we calculated two partial perplexities in addition to the total perplexity. The first P_1 refers to all words with zero counts ($N(w) = 0$), the second, P_2, covers the words with $N(w) \neq 0$. The results of the category model without and with a cache, based on 119 (German corpus) and 149 (English corpus) automatically produced categories, can be found in Table 1. A cache length of 30 was chosen.

Table 1: Test set perplexities of a category model with and without a cache

category model	German corpus			English corpus		
	P_1	P_2	P	P_1	P_2	P
without cache	64834	287	553	840646	321	513
with cache	55821	273	520	184155	267	401

Comparing the perplexities of both language models we see that the most obvious result is the difference in decrease between the German (6%) and the English (22%) corpus. This is caused by the greater heterogeneity of the English texts. The short-term frequencies differ much from the long-term frequencies of the training data, so that the adaptation has a greater effect. Another interesting fact is the considerable difference between the amount of reduction of the two partial perplexities. The explanation is that in the case of P_2 the word probabilities are estimated more reliably for the words that were observed at least once in training, so that the gain of the cache part is not as big as it is for P_1. For the same reason, the loss caused by the critical events with $N_k^c(w) = 0$ is compensated by the word probabilities.

This shows that the pure adaptation effect is measured by P_2 whereas P_1 is mainly affected by an improved handling of words not observed in the training text. If those words occur in bursts, the cache model starts improving their probability just after the first observation. As P_1 covers only 12% (German text) and 4% (English text) of the respective testing data, the amount of improvement is mainly determined by P_2.

Now let us turn to the second variant. The cache-word model requires the addition of only one cache containing the last n predecessor words of the word to be predicted. Since it reflects an immediate temporal frequency distribution it can be used to update the unigram probabilities:

$$q^c(w) = \beta\, q(w) + (1-\beta)\,\frac{N^c(w)}{N^c},$$

where $N^c(w)$ is the cache frequency of word w and N^c the total number of cache elements.

The cache causes the same reduction tendencies in the word model as it does in the category model (Table 2). With a cache length of 300 and 1000 respectively we attained an overall improvement of the same size, 7% for the German corpus and 21% for the English one.

Table 2: Test set perplexities of a word model with and without a cache

word model	German corpus			English corpus		
	P_1	P_2	P	P_1	P_2	P
without cache	177860	302	653	3745332	322	543
with cache	147267	285	607	675254	272	427

In summary we can say that the cache model provides a good method for handling sparse data and enables the language model to adapt itself to changing topics in the test set. It is not important whether the cache is used in a word or a category model. The improvement depends only on the amount of words with zero counts in the training set and on the heterogeneity of the corpus.

[Kuhn 1990] R. Kuhn, R. de Mori: "A Cache-Based Natural Language Model for Speech Recognition", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 12, pp.570-583, June 1990.

August Fenk und Gertraud Fenk-Oczlon

DAS MENZERATHSCHE GESETZ UND DAS PRINZIP DES KONSTANTEN INFORMATIONS-FLUSSES

Das sogenannte "Menzerathsche Gesetz" basiert vor allem auf der folgenden, von MENZERATH (1954) am Deutschen beobachteten Regularität: Die Phonemzahl pro Silbe korreliert negativ mit der Silbenzahl pro Wort. Inzwischen vertritt eine wachsende Zahl überwiegend theoretisch orientierter Analysen (z.B. ALTMANN 1980, GERLACH 1982, GROTJAHN 1982, KÖHLER 1982, HEUPS 1983) die Auffassung, daß die Gültigkeit "dieses" Gesetzes nicht auf das Wort-Silben-Verhältnis und/oder nicht auf das Deutsche beschränkt sei; für natürliche Sprachen könnte ganz allgemein und im Sinne einer Sprachuniversalie gelten: "The longer a language construct the shorter its components (constituents)" (ALTMANN 1980, S. 2).

So allgemein formuliert und verstanden läßt sich das Menzerathsche Gesetz auch auf Befunde "cross-linguistischer" Untersuchungen anwenden, in denen die Beziehung zwischen den jeweils interessierenden Dimensionen (z.B. Satzlänge/Silbenlänge) über die entsprechenden statistischen Kennwerte einer größeren Anzahl von Einzelsprachen ermittelt wird. Angesichts solcher Befunde leistet das verallgemeinerte Menzerathsche Gesetz zweierlei:

(1)
Erstens erlaubt es die Erklärung einer experimentell festgestellten, systematischen Beziehung zwischen der Satzlänge in Silben und der Silbenlänge in Phonemen: Die von 26 native speakers in ihre jeweiligen, typologisch sehr unterschiedlichen Muttersprachen übersetzten deutschen "Kernsätze" (n = 22) variierten zwar in ihrer Länge nur innerhalb des engen Bereichs von 7 plus minus 2 Silben (FENK--OCZLON 1983); aber die Variation innerhalb dieses Bereiches zeigte ein System (FENK-OCZLON & FENK 1985), welches unschwer unter dem verallgemeinerten Menzerathschen Gesetz zur Deckung zu bringen ist: Je größer die Anzahl der Silben, welche eine Sprache zur Codierung einer Bedeutungseinheit bzw. zum Paraphrasieren eines deutschen Kernsatzes benötigt, umso geringer ist die Silbenkomplexität in diesen Sprachen (r = -0,77, sign. 1%).

(2)
Zweitens ermöglicht es die Aufklärung eines Widerspruchs zwischen einerseits der von FUCKS (1956; 1964, S. 15) vorgeschlagenen, über verschiedene Sprachen hinweg nicht-linearen Funktion zwischen Wortlänge (in Silben) und Wortinformation (in dit), und andererseits der kognitionstheoretisch plausiblen Annahme einer - Linearität implizierenden - Proportionalität zwischen Informationsmenge und Verarbeitungszeit (FENK & FENK 1980), wonach Wörter mit größerer Silbenzahl im Mittel ein Mehr an Information transportieren und einen im selben Verhältnis vergrößerten Zeitaufwand (zur Produktion und Perzeption) benötigen. Aus diesem Blickwinkel des verallgemeinerten Menzerathschen Gesetzes entpuppt sich dieser Widerspruch als ein scheinbarer: Wenn längere Wörter häufiger aus eher kürzeren Silben zusammengesetzt sind, und wenn die pro Zeiteinheit bewältigte Informationsmenge "konstant" ist, dann kann die Beziehung zwischen der Länge und dem Informationsgehalt der Wörter nicht linear sein.

(1) und (2) dürfen als empirische Argumente für eine Verallgemeinerung des Menzerathschen Gesetzes gelten. Anders gesagt: Im

Sinne der von COOMBS (1984) getroffenen Unterscheidung zweier Dimensionen erfahrungswissenschaftlichen Fortschrittes - "generality" und "power" - kann von einem erfolgreichen Versuch gesprochen werden, das Menzerathsche Gesetz zu generalisieren bzw. seine Domäne auf eine neue Kategorie von Erfahrungsinstanzen auszudehnen.

Wenn es aber gilt, das Menzerathsche Gesetz seinerseits zum Gegenstand von Erklärungsversuchen zu machen, ist wohl am ehesten an kognitive Ökonomieprinzipien zu denken:
Die psychische Präsenzzeit ("Gegenwartsdauer") könnte auf der Ebene der Satzverarbeitung maßgeblich sein und von dort her auch in die Wort- und Silben-Ebene hineinwirken: Wörter können nicht aus beliebig vielen Silben zusammengesetzt sein, sollen sie innerhalb eines nicht beliebig langen (Teil-)Satzes auf verschiedene Weisen untereinander kombinierbar sein. Dasselbe gilt dann sinngemäß für Wörter als Bausteine von Komposita, für Silben als Bausteine mehrsilbiger Wörter, und schließlich auch für Phoneme als Bausteine komplexerer Silben. Mit zunehmender Verwendungshäufigkeit werden komplexere Gebilde erstens geläufiger und zweitens - auf Kosten der "Transparenz" - "abgeschliffen" und sparsamer kodiert; gerade jene Einheiten, die bereits geläufig und kurz geworden sind, bieten sich als Bausteine eines neuen, komplexeren Gebildes an, dessen Verarbeitung unsere Kapazität - auf der jeweiligen (Silben-, Wort-, Satz-)Ebene der Verarbeitung - nicht überfordert (FENK & FENK 1987). Dieses im Grunde genommen einfache Modell skizziert einen Rahmen, innerhalb dessen sich die typologische Ausdifferenzierung - z.B. hinsichtlich Silbenkomplexität und Morphosyntax - eben nicht beliebig, sondern systemerhaltend, "systemangemessen" (WURZEL 1984) und "selbstregulierend" (KÖHLER 1986) vollzieht.

Wir haben das Menzerathsche Gesetz also einmal als erklärendes und dann als zu erklärendes Prinzip betrachtet. Alle dabei ins Treffen geführten Argumente laufen darauf hinaus, daß das Menzerathsche Gesetz gewissermaßen im Dienste der Erhaltung eines "konstanten" sprachlichen Informationsflusses steht.

LITERATUR

Altmann, G. (1980): "Prolegomena to Menzerath's Law", in: Grotjahn, R. (Ed.): Glottometrika 2. Bochum (Brockmeyer), 1-10.

Coombs, C.H. (1984): "Theory and experiment in psychology", in: Pawlik, K. (Ed.): Forschritte der Experimentalpsychologie. Berlin/-Heidelberg (Springer), 20-30.

Fenk, A., G. Fenk (1980): "Konstanz im Kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß?", in: Zeitschrift für experimentelle und angewandte Psychologie 27, 400-414.

Fenk, A., G. Fenk-Oczlon (1987): "Semiotics and the logic of empirical research in naturalness theory", Beitrag zum XVth Congress of linguists, Abstracts: 452. Berlin.

Fenk-Oczlon, G., A. Fenk (1985): "The mean length of propositions is seven plus minus two syllables - but the position of languages within this range is not accidental", in: d'Ydewalle, G. (Ed.): Cognition, information processing, and motivation. Amsterdam (Elsevier Science Publishers B.V.), 355-359.

Fucks, W. (1956/1964): "Die mathematischen Gesetze der Bildung

von Sprachelementen aus ihren Bestandteilen", in: Nachrichtentechnische Fachberichte, Beiheft der NTZ, Bd. 3, 7-21.

Gerlach, R. (1982): "Zur Überprüfung des Menzerath'schen Gesetzes im Bereich der Morphologie", in: Lehfeldt, W., U. Strauss (Eds.): Glottometrika 4. Bochum (Brockmeyer), 95-102.

Grotjahn, R. (1982): "Ein statistisches Modell für die Verteilung der Wortlänge", in: Zeitschrift für Sprachwissenschaft, 44-75.

Heups, G. (1983): "Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen", in: Köhler, R., J. Boy (Eds.): Glottometrika 5. Bochum (Brockmeyer), 113-133.

Köhler, R. (1982): "Das Menzerathsche Gesetz auf Satzebene", in: Lehfeldt, W., U. Strauss (Eds.): Glottometrika 4. Bochum (Brockmeyer), 103-113.

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

Wurzel, W.U. (1984): Flexionsmorphologie und Natürlichkeit. Ein Beitrag zur morphologischen Theoriebildung. Berlin: Studia Grammatica 21.

Hans Goebl

DIALECTOMETRY: A SHORT OVERVIEW OF THE PRINCIPLES AND THE PRACTICE OF QUANTITATIVE CLASSIFICATION OF LINGUISTIC ATLAS DATA

1. Dialectometry as quantitative classification of linguistic atlas data.

   Since its explicit founding in 1971 by Jean Séguy, Professor of Romance linguistics at Toulouse, dialcotometry has always been defined as a method of numerical synthesis of a possible maximum of linguistic atlas data. The scope of this global data synthesis - an inductive method proceeding from the particular to the general - has been the formation of heuristically useful classes.

   Being a taxometric discipline, thus a procedure of numerical classification, dialectometry shares both research problems and objectives with taxometric branches of many other Human and Natural sciences: measurement of raw data, setting up of a data matrix, choice of distance or similarity indices, setting up of a distance or a similarity matrix, univariate or multivariate processing of the matrices, making up of impressive cartographic visualizations of the data output serving (geo)linguistic questioning in a constructive way (choropleth and isogloss maps, stereograms, etc.), interpreting the results of numerical data synthesis, setting up of a stimulation design of further dialectometric research, etc.

2. "Extra atlantes linguisticos nulla salus dialectometrica".

   Dialectometry entirely depends on its data base, the linguistic atlas. By adjoining and completing (but not competing with!) traditional means of qualitative analysis it is an original method of reading linguistic atlas data synthetically and no longer monothetically, i.e. interpreting only isolated items.

   It seems desirable - starting from the existing dialectometric experiences (regarding some Romance, English and German linguistic atlasses for the most part) - to gather and to carry on dialectometric experiences on the basis of the largest possible number of thematically varying linguistic atlasses.

3. Dialectometry: a challenge to the cartographic competence of geolinguists.

   The concepts of "linguistic atlas" and "linguistic geography" were introduced into geolinguistics quite a long time ago, a series of cartographic means of data analysis (isogloss maps, mute maps, etc.) having been developed since then. Theses methods, however, appear to be highly insufficient, if one considers those already existing in traditional cartography on one hand, and those nowadays being elaborated in computer aided cartography on the other. Furthermore, a methodically reliable treatment of complex cartographic material promotes a respectable improvement of the theoretical level of geolinguistics by rising new questions and presenting correspondingly unconventional answers.

4. Dialectometry: a key to interdisciplinary cooperation.

   Traditional analyses of linguistic atlasses have already clearly given evidence of a close affinity of human linguistic (dialectal) management of a given area with a number of social, anthropological and economical management modalities of the same space.

   And it is now possible to analyse all these modalities in the same taxometric way - giving us an overall insight in spatial data of various kind - a considerable amount of very interesting and methodically safe interdisciplinary correlations will allow to newly define our theoretical concept of the interaction between humans and space.

Rüdiger Grotjahn

MESSEN UND STATISTISCHE DATENANALYSE IN DER LINGUISTIK

Die Beurteilung der Angemessenheit von Meßmodellen und statistischen Analysemodellen in der Linguistik hängt in starkem Maße von impliziten und/oder expliziten ontologischen Grundannahmen über die Beschaffenheit des jeweiligen Erkenntnisobjekts ab. So kann Sprache z.B. als ein System von Symbolen mit bestimmten syntaktischen, semantischen und phonologischen Eigenschaften, aber auch als spezifisches Mittel sozialer Interaktion angesehen werden. Im ersten Teil des Vortrages soll kurz auf unterschiedliche ontologische Annahmen über Sprache eingegangen werden und einige generelle Beziehungen zwischen den diskutierten Annahmen und möglichen Meß- und Analysemodellen aufgezeigt werden.

Daran anschließend soll näher auf den Meßbegriff eingegangen werden. Messen wird im weiteren expliziert als eine spezifische Art der Modellbildung, nämlich als Modellbildung mit Hilfe numerischer Systeme. Es wird dabei mit Gigerenzer (1981) von einer fünfstelligen Modellrelation M(S, O, E, N, Z) ausgegangen, wobei S das Modellsubjekt, O das Modelloriginal, E das empirische Relativ, N das numerische Relativ und Z die Zielsetzung der Messung bezeichnet. Zentral für die dargelegte Konzeption ist die These, daß die Anwendung eines numerischen Systems auf den Gegenstand Sprache stets eine linguistische Theorie über diesen Gegenstand impliziert. Die vorgestellte Meßkonzeption wendet sich damit entschieden gegen die häufig anzutreffende Auffassung, daß numerische Systeme und Methoden (z.B. Skalierungsverfahren, Korrelationsrechnung) eine bloße Werkzeugfunktion haben. Die Anwendung von Meßmodellen und statistischen Verfahren ist vielmehr als ein theoriegeleiteter Eingriff in den untersuchten Gegenstandsbereich anzusehen. Erst im Lichte von Theorien bekommen Meßwerte und Ergebnisse statistischer Methoden einen Sinn.

Auf dem Hintergrund dieser Ausführungen werden exemplarisch folgende drei Problembereiche diskutiert: (1) Indexbildung; (2) Signalentdeckungstheorie; (3) Messung von sprachlicher Variabilität mithilfe variabler Regeln und ableitungsbewertenden Grammatiken.

(1) Indexbildung. In der quantitativen Linguistik wird sehr häufig anhand von Indizes gemessen. Es können zwei diametral entgegengesetzte Typen von Indizes unterschieden werden. Zum einen kann im Sinne der Repräsentationstheorie der Messung die Beziehung zwischen Index und Merkmal auch aus einer bloßen Definition bestehen. Im letzteren Fall liegt eine willkürliche Messung oder Messung per fiat vor. Die weitaus größte Zahl der in der quantitativen Linguistik publizierten Indizes beruhen auf einer Messung per fiat. In vielen Fällen weiß man nichts oder nur sehr wenig über deren Reliabilität und Validität. Auch die statistischen Eigenschaften (z.B. Stichprobenverteilung) sind meist nicht bekannt.

(2) Signalentdeckungstheorie (ST). Die ST ist ein z.B. in Hinblick auf die Semiotik besonders interessantes Skalierungsmodell. Die ST erlaubt die Definition unterschiedlicher Typen von idealen Empfängern (Beobachtern, Hörern, etc.) und die Analyse beobachteten Verhaltens in Hinblick auf das jeweils definierte Ideal. Unter der Annahme, daß das interne (sensorische) Kontinuum einer Versuchsperson (Vp) z.B. einer Normalverteilung gehorcht, wird anhand des tatsächlich beobachteten Entscheidungsverhaltens der Vpn bei der Präsentation von Reizen auf die kognitiv-sensorische Leistungsfähigkeit der Vpn geschlossen. Das psychophysische Modell der Signalentdeckungstheorie erlaubt damit eine datengestützte Modellierung kognitiver und sensorischer Entscheidungsprozesse z.B. bei metalinguistischen Akzeptabilitätsurteilen.

(3) Messung sprachlicher Variabilität mithilfe des logistischen Modells variabler Regeln und mithilfe ableitungsbewertender Grammatiken. Ableitungsbewertende Grammatiken sind Beispiele für sog. probabilistische Parametermodelle, die wiederum einen Spezialfall abgeleiteter Messung darstellen. Es soll u.a. auf folgende Fragen eingegangen werden: Welche Relationen im empirischen Relativ werden durch das Meßmodell überhaupt abgebildet? Handelt es sich bei den gemessenen Wahrscheinlichkeiten um objektive Wahrscheinlichkeiten oder eher um subjektive Wahrscheinlichkeiten, die zudem entscheidungstheoretisch interpretierbar sind? Messen variable Regeln lediglich die statistische Variabilität in einem bestimmten Korpus, oder messen sie eine probabilistisch zu interpretierende Kompetenz?

Im zweiten Teil des Vortrags sollen insbesondere folgende Problembereiche diskutiert werden:

(1) Vorteile von quantitativen gegenüber qualitativen Begriffen. Hier wird z.B. auf folgende in der Literatur genannten Vorteile quantitativer Begriffe eingegangen: (a) Reduktion des wissenschaftlichen Wörterbuchs; (b) Exaktheit der Beschreibung; (c) Exaktheit von Prognosen und Erklärungen.

(2) Hypothesenprüfung und Inferenz. Hier ist eine kritische Auseinandersetzung mit den üblicherweise in der quantitativen Linguistik verwendeten inferenzstatistischen Verfahren vorgesehen. Die Kritik wird sich vor allem auf die weitgehende Vernachlässigung des Fehlers 2. Art sowie auf die ungenügende Berücksichtigung von Effektmaßen beziehen.

(3) Deduktion und Induktion bei der linguistischen Modellbildung. Hier wird u.a. auf die Vorteile von Differential- und Differenzgleichungen bei der deduktiven linguistischen Modellbildung eingegangen sowie auf eher heuristische Verfahren wie die nicht-konfirmatorische Faktorenanalyse. Es wird u.a. dargelegt, daß selbst bei einem optimal angepaßten Modell bei steigendem Stichprobenumfang die Wahrscheinlichkeit wächst, daß ein statistischer Modelltest (z.B. ein Chi2-Anpassungstest) gegen das jeweilige Modell spricht. Aus diesem Grunde ist eine Entscheidung über Annahme und Ablehnung im Fall eines einzelnen statistischen Modells prinzipiell problematisch. Das geschilderte Problem stellt sich in besonderem Maße in der quantitativen Linguistik, in der häufig mit sehr großen Stichprobenumfängen gearbeitet wird. Es wird dafür plädiert, für ein und denselben Datensatz zwei oder mehr konkurrierende Modelle zu entwickeln und für jedes Modell die Anpassungswahrscheinlichkeit zu berechnen und sich erst dann für ein bestimmtes Modell zu entscheiden.

Literaturhinweise:

Altmann, G., R. Grotjahn (1988): "Linguistische Meßverfahren", in: Ammon, U., N. Dittmar, K.J.Mattheier, K. J. (Eds.): Sociolinguistics/Soziolinguistik: An International Handbook of the Science of Language and Society. Berlin, 1026-1039.

Gigerenzer, G.(1981): Messung und Modellbildung in der Psychologie. München.

Grotjahn, R. (im Druck): "Daten und Hypothesen in der Semiotik", In: Posner, R., K. Robering, Th.A. Sebeok (Eds.): Semiotics. A Handbook of the Sign-theoretic Foundations of Nature and Culture. Berlin.

Rolf Hammerl

EIN MODELL ZUR BESCHREIBUNG MULTIVARIATER RELATIONEN LEXIKALISCHER EINHEITEN

1. Methodologische und methodische Grundlagen der synergetischen Linguistik
   - Sprache als selbstregulierendes System; Zusammenwirken von Bedürfnissen, Ordnungsparametern, funktionalen Äquivalenten
   - Synergetische Forschungsmethode: Hypothesen - Modellierung
   - empirische Überprüfung - theoretische Validierung
   - drei Gruppen von Systembedürfnissen nach R. Köhler
   - allgemeiner Modellierungsansatz von R. Köhler und G. Altmann

2. Lexikalisches Basismodell von R. Hammerl
   - Untersuchungseigenschaften: Länge, Häufigkeit, Polylexie
   - Systembedürfnisse und Ordnungsparameter des Basismodells
   - mathematische Modellierung von Wahrscheinlichkeitsverteilungen am Beispiel des Krylovgesetzes: Weber-Fechnersches-Gesetz - modifizierte Zipfsche Verteilung
   - mathematische Modellierung von zweidimensionalen funktionalen Abhängigkeiten am Beispiel der Abhängigkeit der Polylexie von der Frequenz: modifizierte Zipfsche Funktion
   - mathematische Modellierung dreidimensionaler funktionaler Abhängigkeiten am Beispiel der Abhängigkeit der Polylexie von der Frequenz und Länge

3. Anwendung der modifizierten Zipfschen Verteilung auf die Beschreibung des Krylovgesetzes
   - Ergebnisse der empirischen Überprüfung der im Kapitel 2 abgeleiteten Verteilung an polnischen Daten aus dem Projekt "Sprachliche Synergetik" und an zusätzlichen Daten zur polnischen, russischen, französischen, portugiesischen und spanischen Sprache

Rolf Hammerl, Jadwiga Sambor

POLNISCHE ARBEITEN IM RAHMEN DES FORSCHUNGSPROJEKTS "SPRACHLICHE SYNERGETIK": MATERIALIEN UND ERGEBNISSE

1. Synergetische Beschreibung der Sprache

1.1. Grundlagen der synergetischen Linguistik

- Sprache als selbstregulierendes System
- Zusammenwirken von Bedürfnissen und funktionalen Äquivalenten
- Sprachgesetze und System von Sprachgesetzen
- induktiv-deduktive Methode bei der Ableitung und Überprüfung von Sprachgesetzen

1.2. Projekt "Sprachliche Synergetik": Ziele

- Erklärung des Regelmechanismus der natürlichen Sprachen, dessen Bestandteile sich in numerischen Relationen ausdrücken lassen
- Aufstellung eines Systems von Sprachgesetzen über die Zusammenhänge zwischen zwei bzw. mehreren Eigenschaften sprachlicher Einheiten und Überprüfung in mehreren Sprachen

2. Modell von R. Köhler als erste synergetische Beschreibung der Lexik

2.1. Untersuchungseinheiten, -methoden und -material
2.2. Elemente des Basismodells: Bedürfnisse und Ordnungsparameter
2.3. Mathematische Modellierung
2.4. Ergebnisse der empirischen Untersuchungen

3. Polnische Version des Forschungsprojekts "Sprachliche Synergetik"

3.1. Materialien

- Grundlage ist das polnische Häufigkeitswörterbuch und zwei einsprachige Wörterbücher, insgesamt 30.041 Lexeme

- vier quantitative Eigenschaften: Länge, Frequenz, Polylexie, Polytextie; Flexionseigenschaften (Wortart, Flexionsparadigma); morphologische Eigenschaften (Typen von Komposita)

3.2. Ergebnisse

3.2.1. Empirische Daten

(a) Verteilung der Wortlängen in Phonemen, Buchstaben, Silben für einzelne Wortarten
(b) Statistik der Wortarten
(c) Statistik der Flexionsparadigmen
(d) Statistik der Kompositatypen

3.2.2. Mathematische Beschreibung und empirische Überprüfung der erhaltenen Modelle für

(a) das "Krylovgesetz" und die Häufigkeitsklassenverteilung
(b) zweidimensionale Regressionsmodelle:

- mittlere Frequenz als Funktion der Länge
- mittlere Polylexie als Funktion der Länge
- mittlere Länge als Funktion der Polylexie
- Polylexie als Funktion der mittleren Länge
- Länge als Funktion der mittleren Frequenz

(c) dreidimensionale Regressionsmodelle

- Polylexie als Funktion der mittleren Frequenz und Länge
- Länge als Funktion der mittleren Polylexie und Frequenz

3.2.6. Globaler Regelkreis des lexikalischen Basismodells der Lexik als Konsequenz der im Kapitel 3.2.5. dargestellten Untersuchungen

3.2.7. Weitere Untersuchungen

(a) Untersuchungen zum Martingesetz; Lexemnetze (Glottometrika 8,9, 10; Sammelband: Definitionsfolgen und Lexemnetze),
(b) Untersuchungen zur Verteilung der Wortarten im Text (Glottometrika 11, 142ff.),
(c) Vergleich der Längenangabe von Lexemen nach der Silbenzahl im Textwörterbuch und im Lexikon (Glottometrika 10, 198ff.),
(d) Untersuchung einer Hypothese zur Kompositabildung: Zusammenhang zwischen der Länge der Grundwörter und der Zahl der mit diesen Grundwörtern gebildeten Komposita (Glottometrika 12, 73ff.).

Luděk Hřebíček, Prague

TEXT AS A CONSTRUCT OF AGGREGATIONS

The quality "to be a construct" and its equivalent "to have constituents" should be ascribed to all linguistic units of all linguistic levels. When a unit is characterized as a construct, it is natural to ascertain its constituents and vice versa. The concepts of 'construct' and 'constituent' were introduced into linguistics in a systematic way by G. Altmann (1980), cf. also G. Altmann / M.H. Schwibbe (1989). The theory, the central point of which represents the Menzerath-Altmann law, contains criteria for seeking and stating constructs and their constituents.
On the other hand, modern textology (or better: text linguistics) long ago began to search structural formants of texts. Our previous attempts were directed at the formants called 'text aggregations', i.e. constructs the constituents of which are sentences of a given text, cf. L. Hřebíček (1990) and (in the press). The aim of the present paper is to ascertain whether text represents a construct in relation to its aggregations in the sense of Menzerath-Altmann's law. Text is here understood as the highest linguistic unit; it can be assumed that text is a construct of constituents located somewhere in human brains.

1. A Review of the Basic Notions

Text, a phenomenon of natural language, is understood here as a continuous unit which is always a finite unit; consequently, it has its beginning and end. Text is a sequence of a certain number of sentences and a certain number of elementary units. As elementary units words (wordforms) are assumed here; in the following analyses word is a realization of a lexical unit or a semantically interpreted lexical unit. Each text increases sentence by sentence. A text containing k sentences contains k-1 continuous sub-texts, all of which have a common first sentence. Each sub-text can be analyzed as a text, and it is irrelevant who or what decides which sentence is the first and which is the last one. Only texts which originated in some natural conditions are objects of our observations.
Aggregations are defined as such constructs occurring in texts the constituents of which are sentences. Text aggregation is a set of sentences of a text, each of which contains a certain elementary unit. By elementary unit we understand either lexical unit or semantically interpreted lexical unit; the following explications operate with lexical units, as we do not want to complicate these explications with the problem of semantic interpretation. In our previously coined terminology the aggregations based on lexical units are 'vehical aggregations'. However, the results presented in the present paper are also valid for the other sort of aggregations, the 'sign aggregations'. It is evident that each aggregation need not be a continuous part of text.
The notion of aggregation is a consequence of the theory described in G. Altmann (1980) and in G. Altmann / M.H. Schwibbe (1989) concerning language constructs and constituents. In this theory the following relation was derived:

$$y = Ax^b \qquad (1)$$

x = the length of a construct
y = the mean length of constituents
A, b = constants

This formula is the formal expression of Menzerath-Altmann's law. With a negative b it represents a mathematical model of the relation generally formulated as follows:

> The longer the language construct, the shorter its constituents.

The empirical validity of this law was verified in linguistics many times and proved on different levels of different languages. On the basis of the relation (1) a general model was constructed by R. Köhler (1986) operating with four variables: length, "Bedeutungspotenz (Polylexie)", frequency and "Kontextgebundenheit (Polytextie)". The author formulated a complete theory with the variables defined as mutual functions in different relations. The theory of text aggregations appears to be an application of Köhler's general theory in text linguistics. In our previous attempts quoted above, the relation (1) was used for the purpose of the analysis of the supra-sentence structures formed by aggregations. Each text is a complicated structure and it is evident that there must be at least one structural level between the level of sentences and the level of text. It is quite natural to expect that if text is a language unit, then its immediate constituents are not sentences - in analogy to words, which are not immediate constituents of sentences having syntactic structure.
Thus it was proved that aggregations are constructs of sentences. The length of aggregations was measured in the number of sentences x, and the mean length of sentences in aggregations y was in accordance with (1). This means that between the observed and the computed values of y no significant difference was ascertained. The results of the investigation of a corpus of texts have supported the statement concerning text aggregations as constructs and sentences as their constituents.
The question arises whether aggregations are constituents of the constructs called 'text'. Menzerath-Altmann's law was used again for the solution of this problem.

2. Variable n/v

Let us introduce the following variables characterizing text:

n = the length of a text in the number of words;
k = the length of a text in the number of sentences;
v = the number of elementary units (e.g. lexical units)
occuring in a text (i.e. the extent of its vocabulary).

Aggregations are sets of sentences, each sentence containing a given lexical unit. Therefore, a text with v lexical units contains v aggregations. When the assertion defining text as a construct of aggregations is correct, the formula (1) holds with x = v.
For the purpose of the following considerations let us formulate the following simplification: In each sentence an arbitrary lexical unit occurs only once. This simplification is not too far from statistic facts occurring in texts observed. If this statement is correct, then it is evident that i-th sentence having the length ni (in the number of words) occurs as an element of ni aggregations. The total number of elements (= sentences) in v aggregations is

$$\sum_i n_i = n \quad .$$

This is the total length of a text. The mean length of aggregations of the text in the number of words is then n/v.
None of the occurences of elementary units can remain aside off the system of aggregations. Therefore, min v equals to the length of the longest sentence of text, i.e. to max n_i. On the other hand, max v = n; if this holds, each word of the text represents the occurence of a new lexical unit, so that each lexical unit has the frequency equal to 1, and thus n = v. Consequently, the interval of v is

$$\max n_i \leq v \leq n \quad . \qquad (2)$$

The question arises whether Menzerath-Altmann's law is valid for these extreme values of v. Let us suppose two extreme cases of texts, one having the length n = 1, and the other one n = v. As the value of y in (1) represents the mean length of constituents, y = n/v, and the length of the construct in the number of constituents is x = v. Consequently, formula (1) can be rewritten as follows:

$$\frac{n}{v} = Av^b \qquad (3)$$

For the extreme text with n = v = 1 is

$$A = 1 \quad . \qquad (4)$$

From (3) it can be deduced that

$$b = \frac{\ln n - \ln A}{\ln v} - 1 \quad , \qquad (5)$$

in which the fraction on its right-hand side is senseless, as it implies division by zero. Thus we obtain b = -1. The negative value of b indicates that in this extreme case Menzerath-Altmann's law holds.
In the second case, in which each elementary unit is represented by a different lexical unit and n = v, from the formula (5) it follows that

$$b = 1 - \frac{\ln A}{\ln n} - 1 = - \frac{\ln A}{\ln n} \quad . \qquad (6)$$

The resulting value is also negative and thus in accordance with the law. The constant b of different texts appears to be a variable falling into the interval

$$\langle -1; -\ln A/\ln n \rangle \; .$$

3. Text as a Dynamic Process

When a given text is characterized by the mean value n/v, a constant is at hand giving no possibility to consider the functional dependence between the length of constituents y = n/v and the length of the construct (= text) x = v. This is impossible within one and the same text. Theoretically one can admit that a set of texts produced under the same conditions can be analyzed for this purpose. However, texts are always influenced by so many uncontrolled random events that the

results cannot be supposed to be reliable. The problem of the evidence of observed y must be solved in a different way.
Let us suppose each text to be a non-finite form; let us take it as an increasing entity. Be it produced or received, text is always a phenomenon increasing from its i-th to (i + 1)-th sentence - with the exception of the last sentence, of course. Variables characterizing text, when they characterize the increasing text, should also be in accordance with the respective laws. In this way constants change into variables.
However, variable n/v apparently is not a satisfactory characteristic for this purpose. While n with the increasing text steadily (or approximately steadily) increases, the increase of v gradually lowers, so that n/v increases and the parameter b of (1) and (3) is thus positive. This is in contradiction with Menzerath-Altmann's law. The reason is evident: aggregations are constituents crossing each other; one sentence is an element of n_i aggregations and these constituents penetrate each other. Besides that, we measure the mean length of constituents in words, but we need to measure them in sentences. Therefore we must use another characteristics of aggregations.

4. Variable v/k

Let us consider the problem in the following way: From the total number of aggregations v some mean proportion belongs to one sentence. This proportion is evidently v/k. The mean value y = v/k should be verified by observation in texts.
The same characteristics can be obtained as follows: The mean length of aggregations is n/v, and mean sentence length is n/k; both values are given in the number of words. We want to know which portion of the mean sentence length belongs to the mean aggregation, or in short: which is the mean sentence per aggregation. This value is expressed by the complex fraction:

$$\frac{\frac{n}{k}}{\frac{n}{v}} = \frac{v}{k} \qquad (7)$$

Consequently, the relation of v to k can be interpreted in two ways, both being meaningful in the textological sense.
How does the value of v/k change with increasing v, that is, when does the text increase? In other words, it must be ascertained whether the equation

$$\frac{v}{k} = Av^b \quad , \qquad (8)$$

which is a consequence of (1) and (7), is in accordance with Menzerath-Altmann's law. In the case of the positive answer to this question, the observed value of the parameter b must be negative.
A corpus of texts was analyzed from the viewpoint of aggregations as constituents. From all the analyzed texts negative values of b were obtained. In Table 1 and Table 2 the results obtained from two texts are presented as examples of these analyses; the first text was in Turkish and the second one in English. The parameters A and b were computed from the observed v/k and v with the help of the method of least squares, cf. G. Altmann (1980, p. 4).

Conclusions

Each text in a natural language represents a construct of aggregations as its constituents in the sense formulated by Menzerath-Altmann's law. According to our opinion, this conclusion is undoubted. Aggregations are one kind of the possible sub-texts of a text, they are generated in the random processes affiliated to the production and reception of texts. It is beyond dispute that texts and aggregations are complicated phenomena, but their explanation becomes simple and lucid when their stochastic character is taken into account.

TABLE 1: Observed v/k in CT

k	x = v	y = v/k	$y_c = Ax^b$
5	23	4.60	4.48
10	39	3.90	4.27
15	60	4.00	4.11
20	79	3.95	4.01
25	107	4.28	3.90
30	126	4.20	3.84
35	143	4.09	3.80
40	160	4.00	3.76
45	173	3.84	3.73
50	181	3.62	3.72
55	192	3.49	3.70
60	204	3.40	3.68
65	220	3.38	3.65
70	249	3.56	3.61
75	273	3.64	3.58
80	291	3.64	3.56
85	299	3.52	3.55
90	313	3.48	3.54
95	333	3.51	3.52
100	352	3.52	3.50
103	364	3.53	3.49

A = 5.95
b = - 0.0904
CT = Cahit Tanyol, Atatürk İlkeleri, in: Y.Nabi (ed.), Atatürkçülük Nedir?, Istanbul 1969, 105-109

TABLE 2: Observed v/k in RE

k	x=v	y=v/k	$y_c=Ax^b$
5	54	10.80	10.87
10	98	9.80	9.99
15	143	9.53	9.48
20	190	9.50	9.11
25	233	9.32	8.85
30	255	8.50	8.74
35	302	8.63	8.53
40	327	8.18	8.44
42	343	8.16	8.38

A = 19.04
b = - 0.1406
RE = Sir James Redhouse (1811-1892), in: Redhouse Yeni Türkçe-İngilizce Sözlük, Istanbul 1968, X-XI

REFERENCES

Altmann, G. (1980): "Prolegomena to Menzerath's law", in: Glottometrika 2 (ed. R. Grotjahn). Bochum, 1-10.

Altmann, G., M.H. Schwibbe (1989): Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Hildesheim, Zürich, New York.

Hřebíček, L. (in the press): Text in communication: supra-sentence structures.

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

Forming Word Classes by Statistical Clustering for Statistical Language Modelling

Reinhard Kneser, Hermann Ney

Philips GmbH Forschungslabor Aachen
P.O.Box 1980, D-5100 Aachen, Germany

1 Introduction

In a speech recognition system the task of the language model is to provide the a priori probability that a word sentence $w_1 \dots w_m$ will occur. Decomposing this into conditional probabilities and approximating long term history by the last k words results in the so-called k-gram model:

$$p(w_1 \dots w_m) = \prod_{i=1}^{m} (p(w_i | w_{i-k-1} \dots w_{i-1})). \tag{1}$$

An approximation of the conditional probabilities can be done by a so-called class model, where each word w belongs to one or more classes g. We first predict a class and then a word belonging to this class:

$$p(w_i | w_{i-k-1} \dots w_{i-1}) = \sum_g p(w_i|g)\, p(g | w_{i-k-1} \dots w_{i-1}). \tag{2}$$

This can be simplified in the special case where a word w may be in one class $g(w)$ only:

$$p(w_i | w_{i-k-1} \dots w_{i-1}) = p(w_i|g(w_i))\, p(g(w_i) | g(w_{i-k-1}) \dots g(w_{i-1})). \tag{3}$$

An estimate of the difficulty of a given text according to some language model is the perplexity defined as

$$PP = p(w_1 \dots w_m)^{-1/m}. \tag{4}$$

It can be interpreted as the average number of words which the language model proposes for each word position in sentence.

The statistical language model has to supply estimates for the probabilities. The maximum likelihood criterion proposes to take the relative counts from a training set. Even for a huge corpus and for the class model there are many k-grams that will not occur in the training, such that the usual probability estimate for those would be 0. Since we do not want to exclude any word sequence, we have to assign some probability > 0 also to unseen events. One way is to smooth the counts with some better known probability (e.g. k-1-gram).

The main advantage of the class models results from the reduction of free parameters. We thus have relatively more material for training and can better estimate the parameters. Further, if classes are built manually, human knowledge gets incorporated into the language model. However, it is only an approximation to the general model.

One way to get word classes is by linguistic experts. This works quite well, but is very arduous, especially when large corpora are needed. An alternative method is to use a clustering algorithm that forms classes automatically by optimizing a suitable statistical criterion.

2 Clustering Algorithm

We want to find a mapping $w \rightarrow g(w)$, that assigns each word to one of M different word clusters. The perplexity on the training data is used as optimizing criterion. As usual we take the negative logarithm,

so that we have to minimize

$$LP = -\sum_{g1,g2} N(g_1,g_2) \ln(N(g_1,g_2)) + 2\sum_g N(g) \ln(N(g)) - \sum_w N(w) \ln(N(w)), \tag{5}$$

where $N(w), N(g), N(g_1,g_2)$ denote the counts of a word, a cluster and a cluster bigram, respectively, in the training corpus.

The following algorithm is suboptimal in that it finds only a local optimum.

Get some initial $g(w)$
Iterate until some convergence criterion is met
Loop over all words w
Loop over all clusters g'
Look how LP changes if we move w from its old cluster $g(w)$ to g'.
Move word w to that cluster g' which gives the lowest LP.

Moving a word w from its old cluster $g(w)$ to a new cluster g' will change only the counts of $g(w)$ and g'. We will therefore get the new LP from the old one by adjusting just these counts in (5). The complexity of this algorithm is hence of the order $M * M * N * I$, where N is the number of words we want to cluster, M is the number of clusters and I the number of iterations.

Obviously LP decreases at every iteration step and is limited by 0. The algorithm will therefore converge and produce a local, but not necessarily a global minimum.

The algorithm works only for a given value of M, so that this algorithm does not give a direct way to find the best number of clusters. One way to solve this question is to divide the training corpus in two parts and then use the first part to find the clusters and optimize the number M of clusters on the second part.

3 Implementation

The algorithm needs some initial clusters $g(w)$. In our experiments we had one separate cluster for each of the the most frequent M-1 words and one single cluster for all remaining words. Since the move of one word to another cluster will influence the clustering of the next words, the order in which we move words might be crucial. We argue that we know most about the frequent words and should cluster them first. Hence we sorted the words in descending order of frequency.

Other initializations and word orders in general do not much affect the final result, however the speed of convergence becomes much slower.

The algorithm also clusters the words in the lexicon, which did not occur at all in the training set. If we put just all these words together into one cluster, we will not have any information about them. It turned out that the words which appear only rarely in our text give a good estimate also for the words which do not appear at all. So it makes sense to put all words with occurrence frequencies less than a threshold to a separate cluster. We obtained best results when we treated all words the same, which occurred less than 4 times.

4 Experimental Results

We made experiments on two corpora. One corpus of 100,000 words contains German newspaper texts, the other consists of 1.1 million words of miscellaneous English texts. Both corpora are labelled with parts of speech (POS) tags. The German corpus is quite homogeneous while the English has a great variety. We took 1/4 of the corpora apart as test set and 3/4 as training set. The vocabulary comprises all words (full forms) from the complete corpus. The size of the vocabulary is 14,000 for the German corpus and 50,000 for the English one.

First we used the above algorithm on the training set to get word clusters for several numbers of clusters M. Then we calculated the perplexity of a class bigram model on the test set using these clusters and compared the results with the labelled POS and the word uni- and bigram results. The probabilities were estimated from the relative counts and smoothed with unigram probabilities.

Table 1: Perplexity on the test set

number of clusters		English	German
1	= word unigram	1138	1185
30		647	699
60		561	638
90		538	610
120		514	**605**
150		500	608
200		486	609
250		479	621
350		**478**	622
700		484	629
49615 14080	= word bigram	541	 650
153 302	= POS	525	 499

Starting with the special case of one cluster, the perplexity drops fast with increasing number of clusters. Depending on the corpus we have a flat minimum at about 350 respectively 120. From there the perplexity increases again slowly toward the word-bigram value. The best clustering result on the English corpus is even better than the results with the part of speech classes. On the much smaller German corpus this is not the case.

Table 2 lists three examples of clusters created by the clustering algorithm. We see that most of the words belong to the same syntactic class, in our example past tense verbs, nouns and adverbs, respectively. Furthermore there are semantical similarities between the words. The majority of the words in cluster 1 are verbs expressing some kind of motion, and cluster 2 contains mainly titles. We also observe words which appear to be in the wrong cluster; e.g. potato, Toulouse-Lautrec, etc. are no titles, and remote, varies, etc. are no adverbs.

Although most of the clusters are more or less intuitively satisfying, there are also clusters which seem meaningless.

Table 2: Most frequent words of some sample clusters from the English corpus

Cluster 1: *went turned sat moved walked ran drove stepped bent climbed jumped slipped leaned swung flashed rushed wandered switched rode staggered hurried sank dashed drifted jerked leaped peered bowed slid slowed strode stumbled swam blinked knelt pinned roared sits slumped struggled ...*

Cluster 2: *Lord President King Queen Prince Captain Earl General nerve Mayor Major Bishop Princess Temple Count Chief Colonel Commander Admiral High Inspector Madame Mere Albert Arnold Marshal Senator Wood Director Master hydrogen Canon Dante potato Eisenhower Reverend Toulouse-Lautrec full-back prosecutor vice ...*

Cluster 3: *quickly apart slowly rapidly quietly shortly sharply steadily remote exclusively softly sadly varies eagerly ranging instantly nervously ranged urgently briskly impatiently hurriedly reluctantly calmly smoothly extensively appreciably boldly coldly motionless ...*

R. Köhler

SYNERGETIC LINGUISTICS

Synergetics is a special type of systems theoretical modelling whose specific characteristic is the treatment of the spontaneous rise and development of structures. The exponents of this interdisciplinary approach in the field of linguistics have shown that synergetics is also compatible with the functional analytic models and explanatory approaches of quantitative linguistics (sensu G. Altmann). It provides concepts which are applicable to the phenomena of self-regulation and self-organization as they are investigated in quantitative linguistics. Like other self-organizing systems, language is characterized by the presence of cooperative and competitive processes which, together with the external forces of biology, psychology, physics, the social system and others, form the dynamics of the system.

Among the central concepts of synergetics, the slaving principle and the order parameters play an important role: A process whose dynamics follow a second process is called 'slaved' by that process. The order parameters are macroscopic entities which determine the behavior of mechanisms at the microscopic level without being represented there themselves. The possibility of forming new structures crucially depends on the existence of fluctuations which are the motor of development. The possible states of the system (the 'modes'), which appear as a result of the relations in the system, are limited by boundary conditions and order parameters - only those modes survive which fit in with these conditions.

We can distinguish synergetics from other systems theoretical approaches which have been applied in linguistics with respect to performance, method, and epistemological demand. In particular, catastrophe theoretical approaches are of a different nature. They enable description and classification of transitions from one equilibrium state to another 'attractor' state; they are, however, not suitable for the construction of models with explanatory power.

The fundamental structural axiom of synergetic linguistics says that language systems possess self-regulating and self-organizing control mechanisms which change the language towards an optimal steady state and an optimal adaptation to its environment (in analogy to biological evolution). The environment of a language consists of the social and cultural systems using it, the individual human beings with their brains, articulatory apparatus, auditory devices, communicative and other social needs, and their language processing and language acquisition devices, the communication channels with their particular physical characteristics, neighbouring languages, and many other factors.

System requirements, which represent the needs of the elements of the environment of the system constitute another class of axioms. There are three types of requirements: (1) requirements which are constitutive for language (the coding requirement and the specification requirement, both of them with the three aspects of descriptive, appellative, and expressive functions), and the application requirement; (2) language forming requirements (e.g. the redundancy requirement and the economy requirement with its aspects of minimization of sign production effort, coding effort, decoding effort, memory effort, inventory size, of globalization and

centralization of meanings, and of invariance / flexibility of the expression-meaning-relation); (3) the requirements of the control level (the requirements of adaptation and stability).

As in biology, the mechanisms of mutation and selection play an important role: In the speech process there is a constant emergence of deviations and variations in language units, of which only a few are able to prevail. In the long run just those variants of elements of a given language will survive, which contribute in some way to its ability to meet the requirements of its environment.

One of the most fundamental problems for functional analytic explanation in synergetic linguistics is the existence of functional equivalents in language, i.e. classes of linguistic means which meet a given requirement. An illustrative example is the following: Every semiotic system has to meet the basic communicative need for coding (and for specifying) meanings by means of expressions. Languages can fulfill this need by means of four different methods (functional equivalents): Meanings can be coded by lexical, syntactical, morphological, or prosodic means. The extent to which a language provides coding means of a particular type is variable over time and differs among languages. Moreover, it has consequences on the behavior of the system and the value of some of its variables. E.g. the more a language tends to use morphological means for specifying meanings of expressions, the stronger becomes the dependency of polylexy (or polysemy) on word length, i.e. the value of the variable T in the corresponding equation PL = A WL^-T increases with the syntheticism of a language.

A synergetic model has to describe the relevant language- external requirements, the functional equivalents which are able to meet these requirements, and the processes in the system, together with the variables and their interactions. Some of the most basic variables on each level of linguistic analysis are inventory size, element size, and frequency. Others are more specific to subsystems of language, e.g. synonymy and polysemy (polylexy). Similarly, some of the processes under consideration are common to several subsystems, such as unification and diversification, others belong to a single subsystem, e.g. lexicalization, specification / generalization, and globalization / centralization.

In order to be empirically testable, the model must provide equations for the functional dependencies among the variables. They are derived from differential equations and stochastic processes which represent the linguistic structure and dynamics of the system.

The synergetic approach provides concepts and modelling techniques that can explain the phenomena of spontaneous production of structure out of chaos, which can be observed in many kinds of systems: the weather (e.g. cloud layers), kinetic patterns in heated liquids, oscillating chemical reactions, laser light, the emergence of life from inanimate matter, etc. Thus, synergetic linguistics can be characterized as the attempt to construct a dynamic theory of language as a self-organizing system far from equilibrium.

Bibliography

Altmann, G. (1987): The levels of linguistic investigation. TL 14: 227-239

Haken, H. (1978): Synergetics. Berlin/Heidelberg/New York (Springer)

Köhler, R. (1985): Zur linguistischen Synergetik. Struktur und Dynamik der Lexik. Bochum (Brockmeyer)

Köhler, R. (1987): Systems Theoretical Linguistics. TL 14: 241-257

Köhler, R. (1990): "Elemente der synergetischen Linguistik". In: Hammerl, R. (ed): Glottometrika 12. Bochum (Brockmeyer), 179-187

Köhler, R.; Altmann, G. (1983): "Systemtheorie und Semiotik". In: Zeitschrift fuer Semiotik 5: 424-431

Jan Králík

PROBABILISTIC SCALING OF TEXTS

This paper will refer to some experiences based on the Prague Linguistic School tradition.

Advanced computer software enables us to perform some effective scaling procedures which order or re-order any numerical data. Thus, problems of scaling in linguistics, such as ordering texts according to quantitative characteristics, or solving problems concerning disputed autorship, became problems of how to establish or choose the indicators which can be relevant for the scaling criteria, as, e.g. chronology, authorship or any general likeliness. Some of the correlations between such criteria and their indicators are already known, as, e.g., the correlation between the polarity of written and spoken texts and the part-of-speech percentual structure of words. This correlation differs from language to language, but it obviously exists and can be applied if we wish to order or separate texts according to the mentioned point of view.

Often, however, the linguist or anybody else has to solve the situation in which criteria are known, but nobody knows which indicators correspond to the point of view in question. In such case, the task of performing scaling is near the task of investigating a black-box. On the other hand, any indicator we choose can be thought as a cut through the multidimensional space, performed along a single axis. And there can exist as many cuts as we wish. It is clear that some cuts also can exist, the result of which will be fully equal each to other, inspite of having been applied on fully unequal objects. To avoid such irrelevancy or insensitivity, as many different indicators as possible have to be introduced. Even such indicators can be used, for which no certain linguistic explanation is known.

The method which is usually applied afterwards deals with a matrix defined by Petrie. It describes coincidences of the same property presence, or coincidences of presence of indicators of such a property. In the next step, then, a matrix defined by Robinson is usually constructed and, finally, this matrix is re-ordered according to the well known theorem by Kendall into so-called Robinsonean form. The symmetric permutation of rows and columns represents the final solution of the scaling problem. All these steps are of a technical/mechanical character and they can be automatized or computerized easily.

The last real and practical trouble consists of the choice of indicators, the sensitivity of which is usually too rough, so that the poor coincidence of such indicators can hardly be adequate to the majority of linguistic tasks. Often, much more detailed quantification can be introduced and not only coincidence of presence/non-presence but even testing of statistical significance of likeliness and differences can be performed. In other words, probabilistic information can be obtained and used to solve the scaling problem.

For such cases, we attempted to formulate a special generalization of the matrix product and both of the Petrienean and Robinsonean property of matrixes and we also re-formulated the Kendall theorem into the following form:

Let X be a row-permuted generally Petrienean matrix. The

row-permutation, which - being applied to matrix X - converts X into the generally Petrienean type, equals the permutation which - being applied to the columns and rows of the symmetric square matrix S = /X,X'/ - converts S into the generally Robinsonean type.

A new proof for this generalized Kendall theorem has been already given and published by the author. The new theorem can deal with very fine quantification of likeliness according to any axis within the mentioned allegorical cut through the black-box. If the values of the testing criteria are used, as already mentioned, e.g., then we can even speak about fully probabilistic scaling.

This theoretical basis enables us not to lose the sensitivity of very detailed quantitative data and to perform scaling and/or clustering of texts in a new way. The more quantitative data is available by computational technique, the more probable or even exact results can be obtained. Examples we tried on material from present-day Czech and from the historical periods of the 19th and 14th centuries confirmed the very good effectiveness of probabilistic scaling of texts as suggested above.

Dr. Jan Králík
Czech Language Institute
Czechoslovak Academy of Sciences
Letenská 4
118 51 Praha 1
Czechoslovakia

Ju. K. Krylov

WAVE THEORY OF COHERENT TEXT GENERATION AND PROSPECTS FOR ITS DEVELOPMENT

1. The idea of extending wave/particle quantum-mechanical dualism to language/speech production process dichotomy put forward in 1983 (Krylov, Jakubovskaia, 1983) has been making some progress since then. The present paper discusses some results in this field (see also Polikarpov, Tuldava, 1989).
   High importance of investigations in the field of text generation wave theory is due not only to problems of theoretical and applied linguistics but also, not to lesser degree, to steadily growing interest in examination of any complex self-organizing systems. In this connection achievements made in applying the formalism of quantum mechanics to description of statistical relationships between units of various levels of speech generation process organization are also of certain interest in terms of general philosophical problems, logic and methodology of generalized quantum theory development.

2. We will consider text generation as a multi-dimensional random process developing in a space having countable number of admissible states. Each state has its eigenfunction $\psi_\kappa(x)$, where x is continuous coordinate (address) of a consecutive occurrence of a running word $x \in [0,N]$, (N - full length of the text). It is natural to assume that all occurrences of words having fixed multiplicity in the text are related to the same k-th state. The probability of k-multiple word belonging precisely to a section of the text from x to x + dx is given by the element of probability $\psi^2(x)dx$. The theory relies on the well-known theorem stating that any quadratically integratable function $\psi(x)$ allows expansion into generalized Fourier's series using an arbitrary ortonormalized system of eigenfunctions $\psi_\kappa(x)$. For a system displaying completeness, coefficients of linear combination

$$\psi(x) = \sum_\kappa c_\kappa \psi_\kappa(x) \qquad (1)$$

suit the equation of closedness:

$$\int_0^N \psi^2(x)\, dx = \sum_\kappa c_\kappa^2 \qquad (2)$$

After evident normalization

$$\int_0^N \psi^2(x)\, dx = 1$$

relation (2) allows to interpret c_κ^2 as probabilities of "activation" of appropriate states and $c_\kappa^2\psi_\kappa^2(x)dx$ as unconditional probabilities of k-multiple words belonging to the interval [x,x+dx]. Now we postulate that all statistical properties of the text are already fixed, if its "wave" function $\psi(x)$ is preselected. It should be stressed that one of the merits of the theory is its openness to construction of various models of text formation differing by the choice of $\psi(x)$ and the specific system of

eigenfunctions $\psi_\kappa(x)$. After the functions are defined within the framework of a specific model, to calculate c_κ it is enough to multiply (1) by $\psi_\kappa(x)$ and to integrate the expression in the interval from 0 to N. As a result, we have

$$c_\kappa = \int_0^N \psi(x)\ \psi_\kappa(x)\ dx \qquad (3)$$

To describe the lexical (frequency) spectrum, it is possible to use eigenfunctions of Hamilton's operator as $\psi_\kappa(x)$

$$\hat{H} = -\frac{\hbar^2}{2} \sum_l \frac{\Delta_l}{m_l} + U(x_1 \ldots x_l \ldots) \qquad (4)$$

where $U(x_1 \ldots x_l \ldots)$ - potential function depending both on the macrostructure of the text being generated and on the character of interaction between constituent elements, Δ_l - well-known Laplace's operator, m_l - individual characteristic of l-th element of the vocabulary, which depends on the choice of units being calculated (word forms, lexemes etc.), $\hbar$ - certain universal linguistic constant. The set of eigenfunctions of operator $\hat{H}$ should suit Schroedinger's equation for stationary states:

$$\hat{H}\ \psi = \lambda\ \psi \qquad (5)$$

3. As a simplest model concretizing the general theory we consider the case of $U(x_1 \ldots x_l \ldots) = 0$. In this case, the system of eigenfunctions is the system of trigonometrical functions $\psi_l = A_l \sin l\pi x/N$. From the conditions of normalization it follows that $A_l = \sqrt{2/N}$. From the assumption of text generation process stationarity it follows that $\Gamma(x) = \text{Const}$. The constant can be easily found from the normalization condition due to which $\Gamma(x) = 1/N$. By using (3) we obtain

$$c_l = \int_0^N \frac{\sqrt{2}}{N} \sin\frac{l\pi x}{N}\ dx = \begin{cases} 0 & l = 2\kappa \\ c_\kappa = \dfrac{2\sqrt{2}}{\pi(2\kappa-1)^2} & l = 2\kappa-1 \end{cases} \qquad (6)$$

As a result, the present model allows to calculate relative frequencies of the lexical spectrum:

$$f_\kappa = c_\kappa^2 = \frac{m_\kappa}{\mathfrak{L}} = \frac{8}{\pi^2(2\kappa - 1)^2} = \frac{B}{(\kappa - 0.5)^2} \qquad (7)$$

where m_l - number of k-multiple words in the text, $\mathfrak{L}$ - its vocabulary. It should be stressed that it was relation (7) which was used by G. K. Zipf as a revised version for approximation of lexical spectrum. Thus, even a simplest version of the theory allows not only to produce formula (7) but also to calculate constant $B = 2/\pi^2$. To compare theory with experiment, relation f_1 was calculated and relevant distribution constructed for 1000 integral texts of lengths ranging from 100 to 10000 words. It emerged that maximum of the distribution falls on the theoretical frequency $f_1 = 8/\pi^2 = 0.81$ which was calculated, as is clear from the above discussion, using not a single adjusting parameter.

4. To give you another example of application of the theory for description of lexical spectra of short texts ($N \leq 1000$ running words) let us consider the model of "potential box" of finite height in which eigenfunctions $\psi_\kappa(x)$ are defined as a set of admissible solutions to Schroedinger's equation:

$$\frac{d^2\psi}{dx^2} + \mu_0[E - U(x)] = 0 \qquad (8)$$

where

$$\mu_0 = m\hbar,\ U(x) = \begin{cases} U_0 & x \leq 0,\ x \geq N \\ 0 & 0 < x < N \end{cases}$$

Here, m and U_0 are parameters accounting: m - for the choice of text vocabulary units being counted and U_0 - for specific properties of the text. It can be shown that in this case frequencies are given by

$$f_\kappa = c_\kappa^2 = \frac{m_\kappa}{\mathfrak{L}} = 2\left(\frac{\eta_\kappa}{\xi_\kappa \mu}\right)\frac{\eta_\kappa}{\eta_\kappa + 1} \qquad (9)$$

The latter model is in effect a one parameter model because it depends only on one resultant parameter

$$\mu = \frac{1}{2}\sqrt{\mu_0 U_0}\ N = \gamma N \qquad \left(\gamma = \frac{1}{2}\sqrt{\mu_0 U_0}\right) \qquad (10)$$

Here, imposing conditions of limiting $\psi_\kappa(x)$ to infinity and of continuity of their first derivatives leaves as admissible only those values of η_κ and ξ_κ which suit the system of transcendent equations:

$$\eta_\kappa^2 + \xi_\kappa^2 = \mu^2 \qquad \eta_\kappa = \xi_\kappa\ tg\ \xi_\kappa \qquad (11)$$

When μ is small, system (11) has the only solution at K = 1 which corresponds to the presence in a very short text of 1-multiple words only. With increasing text length $\mu = \gamma N$ goes up. Solutions for K = 2,3,4... come into being successively (i.e., 2-multiple, 3-multiple etc. words appear in the text). In the extreme case $\mu \to \infty$ the spectrum asymptotically tends to form (7). Thus, the present model allows not only to theoretically calculate spectra of short texts but also to describe change dynamics for these spectra.

One more interesting consequence of the model is that it predicts change character of spectra when using various vocabulary units. Thus, it follows from (10) that one and the same μ can be obtained at different μ_0 provided lengths of the texts are linked by the appropriate relation. Accordingly, if self-similar parameter μ retains the same value, spectrum of relative frequencies of one text calculated, e.g., in terms of word forms, should coincide with spectrum of lexemes of another text. Examination of empirical material at author's

disposal shows that the above effect is really observed. The same effect seems to account for similarity of spectra for texts written in various natural languages provided their lengths suit (10).

5. Assessing the prospects for development of the proposed theory we would like to point out first and foremost that it allows in principle to account for the influence of various levels of speech generation process organization on statistical characteristics of the process by selecting an appropriate form of potential functions $U(x_1 \ldots x_l \ldots)$. Preliminary results obtained in this respect indicate possible emergence of texts whose spectrum is not decreasing monotonously, but features local maxima at some sufficiently large K. Comparison of theory with experiment has shown that such spectra really exist.

   Of certain interest are also analysis and interpretation of continuous spectrum of solutions observed with low probability at $\omega > \sqrt{\mu_0 U_0}$. Existence of these solutions shows that there always exists non-zero probability of generation of texts having "abnormal" spectra (if desired the author can to some extent ignore systemic relationships and produce a text having arbitrarily organized repetitions). Further investigation is required of meaningful interpretation of wave function $\psi(x)$ which within the framework of the above stationary models $\psi(x) \equiv$ Const does not yet permit an unambiguous interpretation. Thus, in addition to interpretation of $\psi^2(x)$ as density of text covering by words of the text's vocabulary, a hypothesis is possible according to which $\psi(x)$ is identified with the autocorrelation function of speech generation process. Analogy with quantum mechanics allows also to assume the possibility of statistical description of text in various mutually complementary representations of $\psi(x)$ when using relevant eigenfunctions $\psi_\kappa(x)$.

   Finally, of substantial interest is further development of the hypothesis (Lebedev, 1983) according to which the nature of the wave process is linked to coding of word images in human brain by packages of neuron activity waves. Note in this connection that transition from the one-dimensional model to the two-dimensional case (round membrans) has demonstrated proximity of high frequency vocabulary spectrum to quantities inversely proportional to natural frequencies of the membrane. In addition, whereas in the one-dimensional case (string) we have a set of frequencies related as terms of harmonic progression: 1, 1/2, 1/3, 1/4, ..., 1/z ..., i.e. classic Zipf's law, in the two-dimensional case the rank distribution is given by

$$F_z = \frac{c}{\lambda_z} \qquad (12)$$

   where λ_z - are roots of Bessel's functions $\Im_n$, arranged in increasing order, i.e. solutions of the equation

$$\Im_n(\lambda_z) = 0 \qquad (13)$$

   Calculations made in accordance with (12), (13) lead to a rank distribution of Zipf-Mandelbrotian type featuring the characteristic bending of the curve (in the zone of small ranks) in log-log coordinates.

REFERENCES

Krylov, Ju.K., M.D. Jakubovskaia (1983): "On Some Possibilities of Applying the Quantum-Mechanical Formalism when Constructing Stochastic Models of Text Generation and Recognition", in: Symposium on Grammars of Analysis and Synthesis and Their Representation in Computational Structures: Summaries. Tallinn, 49-51.

Polikarpov A., J. Tuldava (1989): Volnovaia Teoriia kak Odno iz Napravlenii Kvantitativnoi Lingvistiki Sviaznogo Teksta (Wave Theory, an Important Trend in Quantitative Linguistic Studies of Coherent Text), in: Quantitative Linguistics and Automatic Text Analysis, Tartu, Acta et Commentationes Universitatis Tartuensis, 162-67.

Lebedev A. (1983): Zakonomernosti Provtoreniia Slov v Reci (Regularities of Repetition of Words in Speech), in: Psichologiceskii Zurnal, Vol. 4, No 5, 11-22.

John S. Nicolis

CHAOTIC DYNAMICS OF LINGUISTIC PROCESSES AT THE SYNTACTICAL, SEMANTIC AND PRAGMATIC LEVELS: ZIPF'S LAW, PATTERN RECOGNITION AND DECISION MAKING UNDER UNCERTAINTY AND CONFLICT

We propose a new selection principle which acting via inhomogenous intermittency in phase space (e.g. via a multifractal strange chaotic attractor - the thalamocortical pacemaker) is responsible for a highly non-linear linguistic filtering process, limiting drastically:

a) at the syntactical level: the grammatically legitimate words;

b) at the semantic level: the "interesting" key features of a pattern;

c) at the pragmatic (game-playing) level: the set of possible "moves"

This is a apriori selection principle - whereas "natural selection" is a aposteriori selection principle acting on phenotypes, e.g. on attractors via shifting control parameters and resulting in generation/elimination/fusion ("crisis") of coexisting attractors - "categories".

Multifractal chaotic dynamics applied to information processing devices may thus be instrumental:

a) In drastically simplifying Hardware without compromising the breath of the functional repertoire (Software).

b) In reducing or compressing memory capacity; instead of storing via "pixelization" of a pattern all details, one limits attention to very few key features (associated with small radii of curvature) coding for disproportionally high amounts of stored information.

So, through an inhomogenous intermittent dynamics in state space we filter out only:

Key words
Key features
Key moves

The principle of natural selction in linguistics involves further scrutinization on those small, a priori selected subsets.

Mark Olsen

QUANTITATIVE LINGUISTICS AND HISTOIRE DES MENTALITÉS: GENDER REPRESENTATION IN THE TRÉSOR DE LA LANGUE FRANCAISE, 1600-1950

Socio-cultural historians of France have an important, if rarely exploited, source for the study of the French language, the Trésor de la langue francaise (TLF), a computer database of 2000 French texts containing 150 million words. The TLF was founded in 1957 as part of a project to compile a dictionary of nineteenth and twentieth century French to rival the Oxford English Dictionary. Its purpose was to facilitate a lexicographical analysis of the French language that would be more comprehensive than previous dictionaries. This corpus ranges in scope from major literary and philosophical works to a selection of technical and scientific texts, including works by nearly one thousand authors and four hundred publishers dating from the early 17th century to the 1950s.

The quantitative approach to textual analysis has rarely been applied to this corpus, but is central to our attempt to systematically analyze common features of language. Recent advances in computer technology, both in hardware and software, permit the researcher to analyze huge datasets in ways that were impossible several years ago. This research includes examining the use of a single word or list of words in a single work, all the works by a selected author, or all the works in the entire database. The measures used in this study examine patterns of clustering of words around the term or terms under examination, frequently referred to as collocates. Analysis of collocation is very useful in determining the general meanings of words where the sheer size of the sample make it simply impossible to examine each occurrence manually. Further, changing meanings can be described in simple tables of commonly associated terms. Quantitative analysis of meaning provides a simple, if rudimentary, method for examining the structures of meaning over time, particularly for very common words.

Changing meanings of very common words are of particular interest to historians of mentalité or collective attitudes, since they tend to be used without conscious reflection of their meaning. Our emphasis on the systematic evaluation of common words is an attempt to isolate patterns of meaning of which the speakers are not necessarily aware. The use of very common words also allows us to examine the unconscious structures of thought, for the associations of very frequently used terms are usually automatic and thought to reflect "common sense". The extent to which language shapes individual thoughts and collective actions by providing a structure in which thoughts and actions must occur is open to debate. There is little doubt, however, that linguistic structures play some, at this point indeterminate, role in the organization of perception and culture.

The encoding of gender relations in language has proven to be well suited to quantitative linguistic analysis. We argue that common words like homme, femme, and sexe, change meanings and associations over the four centuries represented in the database. It would seem that gender marking in language changes significantly over time; and that underlying important long-term changes, there are stable linguistic structures in gender representation which suggest that the French language was a clear reflection of the superior political and social position of males in a patriarchical society, until at least the middle of this century.

A number of examples can be cited to demonstrate these conclusions. There are several constellations of meaning that do not change significantly from 1600 to 1950. The category of sexe is not gender neutral. Rather, the words collocated with sexe typically denote the feminine. Among the most frequent and highly correlated words (by measure of standard deviation) are femme, femmes, personnes, beau, faibles (and the older form, foible), faiblesse, and féminin. By contrast, very few of the terms that are frequently correlated with sexe directly refer to the male. Hommes is only moderately correlated with sexe with a standard deviation of 7.8 whereas femmes is the fifth most frequent collocate with a standard deviation of 37.0. The term homme does not appear as a collocate of sexe, since it is collocated with sexe only 160 times (out of 152,210 occurences in the database) and only weakly correlated with a standard deviation of 3.4. What is sexed is female, which is depicted as a weak deviation from the male norm.

The female is defined in French from 1600 to 1950 as a function of the male. The most significantly collocated term with femme in each fifty year period in the years under examination is sa, suggesting that woman is defined as possession by the male. The slippage between the French union of woman and wife/lover/girl friend in femme places the female in a linguistic context of dependency.

Long-term continuities of the linguistic construction of gender is matched by equally important changes in attitudes. The female as object of desire becomes a central element of the representation of the feminine by the end of the nineteenth century. This is, however, not indicated in the TLF sample from 1600 to 1799. Terms denoting emotive relations, such as amour, aimer, and amante are not significantly related to femme in the first two centuries covered by the sample, but become strongly related in the 19th and 20th centuries. The term amour, for example, is negatively related to femme in the 17th century, but becomes very strongly correlated by the 20th century.

Other changes in the usage of femme are just as important, marking the development of new means of differentiating women. In the texts from the first half of the 17th century, one of the most important collocates of femme is mary. The collocation of femme and mari (or the older mary) declines significantly over the period. Several terms replace mari among important collocates several hundred years later, most notably references to the age of women. In 1900-49, the words jeune and vieille ranked 2nd and 4th respectably as the most significantly collocated words. In 1600-49, the words jeune ranked 107th and vieille 141st. By the mid-nineteenth century, age distinctions had become more frequent than any other categorization, including moral (such as honnete, méchante), physical (jolie), or other attributes (charmante). Youth, in particular, seems to be a category of judgment "discovered" in the nineteenth century, as suggested by its jump in rank from 48th to 6th between 1750 and 1850. The development of age attributes as the most frequent means of characterizing women, and the relative decline of mari, suggests that femme became more of an object outside of the confines of marriage, but that this shift encodes the opposition of young/desirable against old/undesirable, an important distinction made from the male perspective. The increased importance of both emotive terms associated with the female and age categorizations reflect a long term cultural revaluation of the feminine, recasting the female in terms of male desire.

Current work on this project can be divided into two distinct areas: development of methodolgy and further substantive research into the relationship of gender, language, and mentalité.

Several methodological improvements are required to make results more significant. The first is refinement of collocation measurements. In our implementation, we have used a simple definition of spans around keywords, measured in either characters or words, without regard to punctuation or syntactic constructions. We have developed, and are in the process of testing, spans of context measured by punctuation marks, allowing us to restrict contexts to sentences and phrases bounded by punctuation marks. We are considering using a light syntactic parser to identify simple syntactic constructions, paricularly for isolation of verbs associated with gendered subjects. There has been little work on span definition for research based on collocations. The statistics used for this study measure the degree to which individual forms are correlated to the keyword using a measure of standard deviation from an expected distribution in subsets of the corpus. We have a prototype morphological analyzer to lemmatize forms in order to consolidate multiple forms of a word into a single table entry. We suspect that results are more scattered than necessary because of the wide range of forms of words.

Substantive research into gender rests on isolation of domain terms to more clearly trace concept clusters in relationship to gender over time. This will consist of selecting a broader range of gender denoting keywords, such as the oppositions implied by monsieur and madame, or fils and fille. A domain dictionary will be established, to identify, for example, terms denoting emotive responses or social power relations. This dictionary will be similar to content analysis dictionaries, but will be used to allow a statistical analysis of broader lexical classes than single words or lemmas. Finally, we are working on creating a clearer linkage to the social history of gender in France during the early modern and modern periods. We have, for example, described gender differences of terms denoting economic and occupational categories. The linkage between social structure and linguistics is not well understood. An examination of gender marked socio-occupational language might give some indication of how discourse and social structure are interrelated.

Quantitative linguistic techniques have been shown to provide insights into the history of mentalités, using the example of gender marking in French from 1600 to 1950. By concentrating on highly frequent words in a large corpus of text, we have attempted to use quantitative methods to understand patterns of expression and association that reflect unconscious attitudes and linguistically coded common sense. The example of gender marked discourse indicates the ways that existing power structures can be embedded into a language, functioning to legitimate and consolidate those power relations.

SEMDAC - Semantic Disambiguation of Verbs and Automatic Classification of Nouns in a Large Corpus of German Texts

Wolf Paprotté

March 7, 1991

1 Project outline

The project aims at the development of tools for the automatic extraction of lexical and conceptual information from a large body of German texts. The automatic classification and concept acquisition will rest on a two-step procedure of statistical analyses linked to robust, rule oriented linguistic parsing of authentic text data resulting in a representation of type sorted semantic objects. The project has a large number of potential applications both for the proposed tools and for the results of the analyses, including handwriting recognition, OCR, disambiguation cues for speech recognition or ambiguous syntactic structures, thesauri, term banks and lexicons, for information retrieval etc.

2 Statistical analyses

Natural languages restrict co-ocurrences of words on the conceptual and the formal syntactic level. In a context which contains *captain*, the words *ship* or *sailor* are more likely to occur than the words*knight* and*castle*. Existing massive bodies of machine readable text data, and facilities for computational storage and analysis have made it possible to generate precise assertions of co-occurrence facts. Using statistical means, the first project step aims at establishing the significant co-occurrences of nouns, verbs and adjectives in our corpus of roughly 50 million text words.

We intend to use the Church / DeRose algorithm(s) (Church 1988; DeRose 1988) to disambiguate word categories; the part of speech information contained in our lexical database will be compared with the results of the statistical analyses. An association ratio algorithm (Church / Hanks 1989) for the extraction of co- occurrences will then be supplemented by two additional parameters for distance measuring, i.e.*spread* (= the distribution of the relative distance between two words) and*height* (= providing a ranking of the strength of spread of two words) (Smadja / McKeown 1990).

The results of the statistical analysis are themselves significant enough to merit support as they will be relevant for OCR, handwriting and speech recognition and furthermore for complex lexicographic tasks, for text and information retrieval and numerous other applications.

3 Rule oriented, computational linguistic analysis

The second step of the project will take as its input the significant co-occurrences and their sentential contexts and submit both to a computational linguistic analysis as to their phrasal types and syntactic functions, parts of which will be mapped into predicate argument structures of various types, e.g. verbal-predicate (subject, object); adjectival-modifier (modified type), etc. A (robust) unification grammatical parser,- we work within the HPSG paradigm (Sag /Pollard 1987; and forthcoming) - in combination with a lexicon of verbal valencies will enable us to extract the semantically richer predicate argument representations underlying statistically significant co-occurrences.

3.1 Nominal classification

The set of arguments of verbal predicates is typically restricted. Groups of nouns sharing co-occurence with a set of verbal predicates are summed over and classified according to their position in the subject or object slot of predicate argument structures. Thus *wine* will collocate with *drink, spill, produce*, but probably not with *cut, fold, break*. Nouns collocating with a similar set of predicates will be good candidates for a class of semantically similar nouns (cf. Hindle 1990). Including the range of adjectival modifiers of nouns as a classification measure will increase the selectivity of the classes of semantically similar nouns.

3.2 Semantic disambiguation of verbal meanings

Both syntactic and semantic parameters will allow us to disambiguate verbal meanings. From a syntactic point of view, number and syntactic function of arguments (syntactic valence) will indicate a subsense of the verb. The co-occurrence of verbs with elements of the same nominal class will increase the likelihood of having identified a verbal subsense.

4 Expected results

We expect the following results and deliverables:

- the development of statistical tools;
- robust linguistic tools for the analysis of authentic linguistic mass data
- the development of a format of semantic representation with frames of predications characterizing conceptual objects and typed conceptual hierarchies;
- the development of a classification of nouns, giving an idea of the "ontology" inherent to a natural language.

5 Prerequisites

The Arbeitsbereich Linguistik at the University of Muenster has accumulated an up-to-date corpus of more than 50 million textwords of German texts which have been produced since 1990; the texts originate mainly with two newspapers, the *Frankfurter Allgemeine Zeitung* and *Die Zeit* but also include technical reports *(VDI Nachrichten)* and a variety of other materials. Texts are still coming in.

Furthermore, a lexical database of some 70 000 entries with complex lexicographic information, a verb valence database, concordance tools, morphologic generators and parsers, and a prototype of a lexicographic workbench (LEXWERK) can be used for purposes of the project.

6 Bibliography

K.W. Church (1988), "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text." *Proceedings of the Second Conference on Applied Natural Language Processing.* Austin, Texas.

K.W. Church; P. Hanks (1989), "Word Association Norms, Mutual Information, and Lexicography." *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics* pp. 76 – 83.

S.J. DeRose (1988), "Grammatical category disambiguation by statistical optimization." *Computational Linguistics* Vol. 14.1 pp. 31 – 39.

D. Hindle (1990), "Noun Classification from Predicate – Argument Structures." *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics* pp. 268 – 275.

Hindle, D.; M. Rooth, (1991), "Structural Ambiguity and Lexical Relations." *Proceedings of the DARPA Workshop on Speech and Natural Language Processing.*

P. Jacobs; U. Zernik (1988), "Acquiring Lexical Knowledge from Text: A Case Study." *Proceedings of the AAAI 88*, St. Paul.

Magerman, D.; M. Marcus (1990), "Parsing a Natural Language Using Mutual Information Statistics." *Proceeding of the AAAI-90.*

R. C. Moore (1989), "Unification -Based Semantic Interpretation." *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics* pp 33 – 41.

S.M. Shieber, G. van Noord, R. C. Moore, F.C.N. Pereira (1989), "A Semantic Head-Driven Generation Algorithm for Unification-Based Formalisms." *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics* pp 7 – 17.

Frank A. Smadja; Kathleen R. McKeown (1990), "Automatically Extracting and Representing Collocations for Langaguage Generation." *Proceedings of the 28th Annual Meeting of the ACL, Pittsburgh 1990* pp. 252 - 259.

K. Sparck Jones (1986), *Synonymy and Semantic Classifi- cation.* Edinburgh University Press.

M. Webster; M. Marcus (1989), "Automatic Acquisition of the Lexical Semantics of Verbs from Sentence Frames." *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics* pp. 177 – 184.

A. A. Polikarpov

ON THE HYPOTHESIS OF WORD LIFE CYCLE

1. The main component of the word life cycle hypothesis is the assumption of semantic potential of a word. Each word emerging in language possesses a certain associative-semantic potential. The potential is manifested by the ability of the word's meanings (images) for associative transitions, interaction with other images presented in the human mind. This versatile ability of the human brain, first of all, provides the basis for linguistic communication giving communicants the possibility for hinting on some senses by means of meanings. Secondly, it is also important for understanding the general possibility of semantic development for each sign and, accordingly, for an understanding of the historical development for language as a whole. The development involves using the mechanism of mutual teaching of communicants to fix, to socialize certain associative transitions, emerging in some communicative acts, from an existing meaning to some extralinguistic sense which is similar to the meaning, to assign to the sense the status of meaning [Polikarpov, 1979].

2. Each sign being originally introduced into language usually for designation of a certain single meaning may consecutively realize its associative-semantic potential in the course of further use by increasing the number of meanings. Realization of the potential means at the same time its spending. For this reason, transition of a word from one polysemantic area to another (i.e. picking up each new meaning) is by no means equiprobable, is steadily slowing down, and the more meanings the word already has, the slower is the process. At some stage the semantic potential becomes exhausted and the development of the word stops.
   Exhaustion of word's semantic development potential, slowing down of the new meaning formation process when moving to higher polysemy zones (till the full stop at some stage) has been confirmed by Russian language data from the "Novye slova i znacenija" (New words and meanings) dictionary [Karimova, Polikarpov, 1988] and English language data from the Supplement to the Oxford English Dictionary [Polikarpov, 1991 - in the press]. Indirect evidence is also available at present for French, Spanish, and Finnish languages.

3. Quite evident premises are also used to forecast the phenomenon of gradual alteration of quality of meanings being consecutively developed by words in the course of semantic development. The main trend of development here is from presentiveness to decreasingly specific feature character, eurysemanticity, and then to syntacticity, specialized eurysemanticity.
   This forecast is preliminarily confirmed by our data on Russian and English languages. However, the need for broad research of vocabulary in terms of this characteristic is evident.

4. One important consequence of qualitative development of further meanings in a word's semantic structure towards increasing abstractiveness, versatility lies in the fact of relatively higher occurence frequency of each new meaning of the word compared to that of meanings developed by the word at earlier stages. Indeed, the less primitive, the less specific, the broader in terms of denotative reference further meanings become, the broader is the range of situations and contexts of communication in which each of them can act as a hinting means. This forecast is confirmed by all available emirical evidence in the form of data on correlation of frequency and polysemy characteristics of words of various languages: Russian, English, Franch, German, Estonian, Finnish, Vietnameses, etc. [Zipf, 1945; Guiraud, 1954; Polikarpov, 1976, 1979, 1991 in the press; Andrukovic, Koroljov, 1977; Koroljov, Korsakova, Safronova, 1984; Tuldava, 1979; Köhler, 1986; Arapov, 1988]. All above sources point to quicker growth of occurence frequency compared to that of polysemy of words, which very fact means that in the course of polysemic development each further meaning of words is on the average more and more common, frequently used.

5. In view of qualitative evolution of consecutively emerging new meanings of a word from presentiveness to feature character and then to syntacticity, the prediction can be made of the most active involvement of meanings of mesosemantic words in synonymic and antonymic relationships. Their feature (qualitative) character should be conducive to this. Meanings of oligosemantic words are too concrete and presentive, meanings of polysemantic words are too abstract and syntactic to become involved in subtle or polar oppositions, which fact is the essence of synonymy and antonymy.
   The above prediction is confirmed by data on Russian and English languages [Polikarpov, 1991a - in the press].

6. Prediction can also be made concerning relative propensity of meanings of words belonging to various polysemic zones for phraseological reinterpretation, i.e. formation of phraseologically bound meanings. The logic of the prediction is here the following: if meanings of mesosemantic words have the most pronounced qualitative feature character, then these are the meanings in which words are most predisposed to various combinations with other words. Some of the combinations may be completely reinterpreted, which very fact means formation of phraseologically bound meanings. Meanings of mono- or oligosemantic words are less versatile and broad to let the words become involved in fixed combinations with some other words.
   Experimental data on the above trends obtained by analysis of the "Novye slova i znacenija" (New words and meanings) dictionary (1984) are reported in [Polikarpov, Meteleva, 1991 - in the press].

7. In addition to the fact of "plus"-development, one should also take into account implications of "minus"-development, i.e. loss of meanings (and, as a result, sometimes loss of whole words) in the course of a word's semantic potential realization. This means that the word life cycle hypothesis must account for certain regular changing at each tact of semantic development probability of original meaning (or other meanings derived from the original one) going out of use. This should also be done on the basis of some general considerations.
   The main assumption here may be the notion according to which earlier meanings of words are, on the average, most active not only in formation of new meanings but also in getting out of use, becoming extinct while later meanings turn

out to be, on the average, less and less subject to this process, featuring steadily declining probability of extinction. This prediction is explained by the same reasons as those used to predict graded positive activity of consecutively emerging meanings. In the early tacts of a word's semantic development, less exhausted potential may be more actively realized also in the opposite direction, the direction of extinction of meanings.
Experimental corroboration of the present prediction still lies ahead. It may be obtained on the basis of comparison of large-size dictionaries of the same language representing different periods of its development. The prediction is not at odds with the well-known fact of the most wide-spread extinction of monosemantic words in any epoch of language existence. However, that fact would also take place in the case where extinction probability of all meanings had been even equal: for words having two, three and more meanings simultanous extinction of several meanings is too unlikely to be too noticeable.
Much stronger corroboration of the supposition of decreasing propensity of subsequent meanings of words for extinction is the fact of high stability of syntactic meanings of words (which emerge precisely at the end of semantic develoment of words).

8. The ability of a word to be involved in homonymic relationships may also be related to its advanced position in the way of realization of semantic potential. However, the picture is rather intricate here because there exist two immediate sources of homonymy: (1) occasional coincidence of sounding (or spelling) of two words as a result of their shortening and elimination of phonetic (or graphical) elements by which they formerly differed and (2) splitting of a formerly whole word into two homonyms as a result of the loss of one link in the chain of consecutive polysemic development and the break-up of links between remaining meanings.
The first source predetermines the dependence of homonymy on the time of existence of the word in language and, accordingly, on its polysemy.
The second source should predetermine the concentration of homonyms in the area of low polysemy because original meanings are most likely to get out of use (see above) which leads to the break-up of the polysemic chain. Data available at the moment corroborates the above prediction [Krylov, Jakubovskaja, 1977, pp. 5-6].
Further consideration is needed for the issue of relative capacity of the two sources of lexical homonymy.

9. Prediction can also be made of increasing survival probability of the word as a whole with each consecutive tact of semantic development because the survival probability of each consecutive meaning is growing and the survival probability of the whole word is tantamount to the logical sum of survival probabilities of all meanings. However, in all likelyhood, this sum of probabilities never becomes equal to 1 because it is known that even syntactic words become extinct from time to time. Qualitative interpretation of a certain extinction probability of syntactic words must account both for inevitable structural, grammatical rearrangements of any language (i.e. periodic systemic elimination of formerly existing grammatical meanings) and for inevitable competition of various synonymic means of expression of the same conceptual category (even on the level of syntactic categories).

If in the course of semantic development a syntactic or autonomous word acquires a meaning in which it becomes involved in a high-frequency syntactic construction (combination of the syntactic word with other syntactic or autonomous words) or in a high-frequency phraseological expression, then the construcion is very likely to undergo step-by-step contraction, the result being the formation of a new lexeme (word form) into which the word is included as a morpheme, a root or an affix.
At a later stage one of the roots of a compound word may also become an affix if the root is productive in formation of compound words in the language and if compound words are sufficiently frequent in speech. The intermediate stage of transformation of a root into an affix became known as semiaffix.
In the course of possible further growth of occurrence rate of the word form, language users may lose or forget the motivation of the word's morphemic composition (e.g., following the loss of independent use of syntactic or presentive words due to competition between words), reinterpretation may occur of a group of morphemes within the word as a single morpheme. This process also leads to further reinterpretation of phonetic envelope of the word form, de-etymologization of its composition, eventual disappearence of any traces of the morpheme which historically has served as a building material for the word. If by that time the prototype word of the affix has really gone out of independent use, then the complete end of the word's life cycle may be stated. For some time the "remnants" of the word (morpheme) have served as a soil for growth of subsequent generations of words, but now they have completely dispersed in cycles of "linguistic metabolism". In the case of complete dispersion of all affixal remnants of the former word in all languages derived from the same parent language and absence of written fixation of stages of their development, etymological analysis of many old "primary" words turns out to be already impossible. However, valuable information on that score is available in the case of languages of some ancient cultures (e.g., Romance languages) whose written monuments reflect some remote states of the language, its intermediate and various modern varieties representing various stages of "digestion" of the original language material.

Paper presented to the First Quantitative Linguistics Conference, Sept. 23-27, 1991, University of Trier, Germany

About Some Theoretical and Computational Interpretations of Chinese Phrase Structure Grammar (CPSG)

Qian Feng

Institut für Computerwissenschaften und Systemanalyse
Universität Salzburg, A-5020, Salzburg, Austria

and

Institut für Computerlinguistik, FB Informatik
Universität Koblenz-Landau
D-5400 Koblenz, Germany

This paper tries to explore some relationship between Chinese Phrase Structure Grammar (CPSG) as a formalism for characterising Chinese language syntax and the applications of it in the computational linguistics (CL). CPSG is loosely based on HPSG (Head-driven Phrase Sructure Grammar, as presented in [Pollard & Sag 1987]) and GPSG (Generalized Phrase Structure Grammar, as presented in [Gazdar et al. 1985]), with certain extensions to capture the peculiarities found in Chinese grammar and an augmentation for the discourse and pragmatic information under the environment of Situation Semantics [Qian 1990(1)], [Qian 1990(2)],[Qian 1991]. The basic concept in CPSG is unification, in this sense, CPSG may be taken as one of what is generally called 'unification-based grammars'.

The auther has been involved in a ongoing research project where the construction of a comprehensive grammar for Chinese and its parser are key elements. CPSG and its parser are being designed for that purpose.

1. Four Properties of Chinese as a Formal Language

Chinese language shows a number of interesting formal properties that are unknown, if not impossible, in Indo-european languages. They deserve hence well-woven consideration in developing a framework for characterising the syntax of this language. The four formal properties following feature Chinese as one of the 'very configurative' languages.

1.1 'Position-orientation'

While Indo-european languages resort inflection to establish the relationship between the sentence itself and the reality, i.e. predication (in the sense of Smirnitzky, see [Qian 1990(1),(2) & 1991]) Chinese uses positioning which proves an assigner to a number of syntactic features. Any phrase, being put into a C (Category) position, will behave as a C under a certain set of constraints.

One of the vital difficulties, however, consists in the fact that there are numerous such different positions and the lexical categorization (parts of speech) in Chinese is not so clear-cut like most Indo-european languages (Otherwise the syntax of Chinese would be rather simple).

1.2 'Sentence-recursion'

Sentences in Chinese language are usually positionally-recursive. To put it another way, there is no syntactic distinction between phrase and clause in Chinese; they show themselves up only relatively in context under certain relevant constraints. In this sense the Indo-european languages are phrase-recursive. Again, however, there seems to be numerous such recursive rules, that makes Chinese syntax sometimes intractable.

1.3 'Knowledge-direction'

This is one of the most remarkable discrepancies between Indo-european languages (also Japanese) and Chinese in syntax. The former features in an independent-syntax, while the latter shows characteristics that the 'incorporated' knowledge plays sometimes a significant role in the syntactic structure of a sentence, for generation as well as for parsing.

1.4 'Logical Organization'

Astonishingly enough, the Chinese version of such sentence as

All that glisters is not gold. (Shakespeare)

is precisely the expression of a first-order predicate culculus :

$$\sim(\forall x)\ (\ G(x) \dashrightarrow Au(x)\ )$$

which represents the correct semantic interpratation of the sentence. Another example brings us to the double negation, like English 'I don't want nothing'. There are a number of tidy logical rules that govern the Chinese double negation that are usually void in Indo-european languages. The seemingly loose syntax of Chinese (as some people still believe) has a rigorous logical flavor.

2. Subcategorization in Chinese language

To cope the formal properties the concept of subcategorization is extended to include categories that are 'higher' than the lexical ones. Basically any category has its own SUBCAT features in CPSG, and this observation works well along with the ID/LP rules in percolating certain features along the phrase structure tree in parsing a sentence.

3. ID/LP rules

3.1 Phrase structure rule

There is only one phrase structure rule in CPSG, namely :

M ---> D H

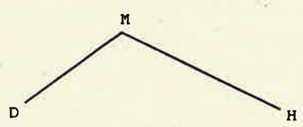
where M denotes the mother node, H, the Head node and D the nonhead daughter node.

3.2 ID/LP rules

The proposed ID/LP rules for CPSG may be presented in the following form :

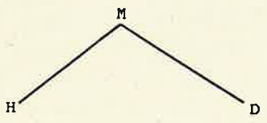
1) A mother with a [SUBCAT <>] dominates a DH pair :

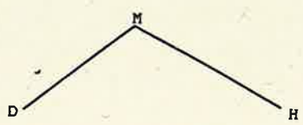
M ---> D H



2) A Mother with a [SUBCAT < k >], where k is any thing other than null, dominates a HD pair unless the D is a [-N]&[-V] :

M ---> H D





D = [-N]&[-V]

4. Unification Conventions and Subcategorization-shift

The unification is defined recursively in CPSG :

1) Two categories X and Y are said to be unified with each other, or X unifies with Y, if for each feature F that has a value defined in both X and Y , so that either of the following holds :

1.1) The value of F in X and the value of F in Y are identical when F is a binary or a multivalued feature;

1.2) The value of F in X unifies with the value of F in Y, when F is a category-valued feature.

2) Two sets of categories are said to be unified, or one of them unifies with the other, if there is a one-to-one correspondance between the two sets and each member of one set unifies with the corresponding member of the other.

3) When unification progresses from a daughter to the mother, and when another daughter is other than [SUBCAT < >], a subcategorization-shift will be automatically evoked. An example of subcategorization-shift is : [SUBCAT < >] ---> np upwards, i.e. in further unification upwards this would be taken as a NP.

5. Extraposition

It seems at first glance confusing with this sub-title. The problem of extraposition is a typical phenomenon in a non-configurative language such as German, rather than in the case of a 'very configurative language' as Chinese. Due to the four properties, especially those with 1.1 and 1.2, extraposition turns a fatal problem also in Chinese garmmar. In CPSG various phenomena of the extraposition are explored and some under traditional Chinese linguistics very difficult problem are tackled. For example, why both of the following :

他在黑板 上 写字。
ta zai heiban-shang xiezi.

字写在黑板上。
zi xie zai heiban-shang. (Extraposition)

are grammatical, but one of the following is not :

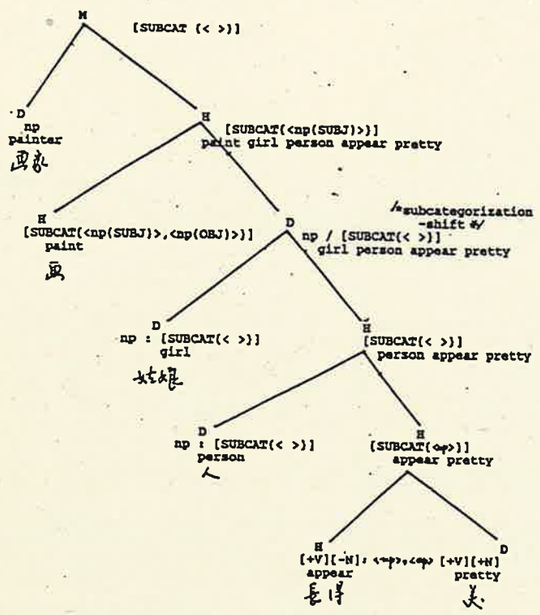
我在台上唱戏。
Wo zai dai-shang changxi.

*戏唱在台上。
* Xi chang zai dai-shang. (Extraposition)

The principles for the construction of extraposition rules, in terms of subcategorization, are presented in CPSG.

6. Analysis of a Chinese sentence in CPSG

画家	画	姑娘	人	長得	美。
Huajia	hua	guniang	ren	zhang-de	mei.
painter	paint	girl	person	appear	pretty



7. On implementation of a CPSG parser in Prolog

There have been several attempts to design GPSG parsers, largely based around Earley's algorithm for parsing context-free languages. Basically they are all active-chart parsing. In designing a ID/LP parser for CPSG we have studied the strategies of Shieber's modifications to Earley's algorithm [Shieber 1984], Kilbury's further modifications to Shieber's parser ('first' relation) [Kilbury 1984], the transitive 'first' relation of Dörre & Momma [Dörre & Momma 1985], and Weisweber's dominance chart parser [Weisweber 1987]. The strategy of the ongoing CPSG-Parser dwells on a kind of dominance chart parsing technique, but with modifications pertinent to the inherent characteristics of Chinese language revailed in CPSG.

In the framework of CPSG, we can visualize the parsing history of a sentence as a process where the values of features of a phrase structure tree are determined by ways of GPSG unification. The unification starts from the information given by each lexical item upwards until the top sign is met. Here unification means the process of proper combination of the ID/LP rules and the principles being applied to the phrase structure tree nodes. The whole structure tree as the result of the analysis is to be constructed step by step with the progress.

References

[Dörre & Momma 1985]Dörre,J.,S.Momma, Modifikation des Earley-Algorithmus und ihre Verwendung für ID/LP-Grammatiken; Manuskript des Institut fUr Linguistik, Universität Stuttgart, 1985

[Gazdar et al. 1985]Gazdar,G.,E.Klein,G.K.Pullum & I.A.Sag, Generalized Phrase Structure Grammar, Oxford : Blackwell,1985

[Kilbury 1984]Kilbury, J., Earley-basierte Algorithmen für direktes Parsen mit ID/LP-Grammatiken, KIT-Report 16; TU Berlin 1984

[Pollard & Sag 1987]Pollard,C. & I.A.Sag, Information-Based Syntax and Semantics, CSLI Lectures Series Notes No. 13. Stanford : CSLI

[Qian 1990(1)] Knowledge-Based CL/MT Systems and Chinese as a Formal Language, The New Medium, 7th Intern. Assoc. for Literary and Linguistics Comput. Conf. and 10th Intern. Conf. on Computers and the Humanities, 4-9 June, 1990 University of Siegen, Germany, pp 78-81

[Qian 1990(2)] A Synopsis of CPSG - Chinese Phrase Structure Grammar, the Intern. Symp. on East Asian Inform. Proc. Oct. 20-21, 1990, University of Pennsylvania, Philadelphia, USA

[Qian 1991] Three Chapters on CPSG, Technical Report, FB Informatik, Universität Koblenz-Landau, Germany (in press)

[Shieber 1984]Shieber, S.M.,Direct Parsing of ID/LP Grammars, Linguistics and Philosophy, 7. 1984, pp 135-154

[Weisweber 1987]Weiweber, W., Ein Dominanz-Chart-Parser für generalisierte Phrasenstrukturgrammatiken, KIT-Report, TU Berlin, 1987

M. Refice, M. Savino

QUANTITATIVE EVALUATION OF LANGUAGE INDEPENDENT MODELS

Simple probabilistic language models have been proved to work successfully in several application tasks.
Whereas real time constraints impose strict limits to the computational complexity, such as in text-to-speech systems, large vocabulary speech recognition sytems, optical character recognizers and so forth, statistical knowledge of the involved linguistic domain may significantly improve the overall performance of such systems.
In this paper we would like to discuss Markovian models based on transition probabilities between pairs and triples of grammatical categories and present several quantitative parameters which are able to measure both the intrinsic power of the models and the quality of the involved statistical knowledge.
Some parts of the described work have been developed within the framework of the Esprit Project "Linguistic analysis of the European languages" which dealt with seven languages, namely Dutch, English, French, German, Greek, Italian and Spanish.
The availibility of similar content domain corpora for all the mentioned languages allowed us to check the validity of the discussed quantitative parameters across the languages, and in some cases a quantitative comparison between languages has also been attempted and will be discussed.
Let's consider, for example, the following simple Markovian model:

$$p(c_k) = p\langle c_k \mid c_{k-1} \rangle$$

This model claims that the probability $p(c_k)$ of finding a grammatical category c_k in a text is simply given by the conditional probability that the category c_{k-1} be followed by the category c_k.
The recognizer may then consist of a simple automaton provided a starting point can be fixed and the set of conditional probabilities for all the category pairs can be known a priori.
As far as the conditional probabilities, which may be called transition probabilities, may be approximated by the corresponding transition frequencies, the latter can be easily computed from a learning tagged text.
The amount of learning needed to make the mentioned approximation, while depending on the quality of the chosen learning text, may be assessed by the information content of the transition frequencies matrix as a function of the text size.
The Entropy, or the log-Probability, as a function of the text size s:

$$LP(s) = \sum_{i=1}^{M} Fc_i \cdot \left(\sum_{j=1}^{M} t_{ij} \cdot \log t_{ij} \right)$$

(in addition to several other considerations not reported in this abstract) may be considered a suitable measurement of the stability of a $|t_{ij}|$ transition frequencies matrix over the total number of categories M, where Fc_i represents the relative frequency of the grammatical category c_i. Our experimental results show that this function is not linear and that a defined limit for LP does exist for reasonable sizes of learning texts for all the corpora taken into account.
Our experiments also show that both the absolute values and the shape of the function depend on the linguistc quality of the chosen text and on the set of grammatical categories; this is especially highlighted when a modified form of the previous formula, considering also the cardinality of the category set, is used.

The mentioned linguistic quality of a text refers quantitatively to the assumed dependence of LP on the text size s.
In linguistic terms, the syntactical structure of a given text modeled by the mentioned transition matrix may be reasonably represented by a relatively small part of it, depending on the specific content domain and its particular style.
If it really occurs, the amount of text needed to let the transition matrix reach its stability may significantly decrease.
In the paper a parameter able to quantitatively characterize the linguistic quality of a text, with respect to the model taken into account, will be presented and discussed. This parameter is given by the following form:

$$LQ = \sum_{i=1}^{C} F_i \cdot \sum_{j=1}^{C} \| d_{ij} / PG_j \|$$

Space limits of the present abstract do not allow a detailed discussion of this parameter; it can be only said that it measures the syntactical constraints of a given text with a total number C of grammatical categories with respect to a text of the same size and total categories, but with the latter randomly chosen one after another.
The results of our experiments will be illustrated by the performances obtained for the mentioned languages by the discussed models.
Some additional considerations will also be given about the relative power of different models.
We have implemented a linguistic recognizer based on a black-board system, which can make use of several models simultaneously in an integrated way along a predefined opportunistic strategy.
The conducted experiments show that a better performance is obtained when different models are cooperating with respect to each single one.
Moreover the possibility will be shown of measuring the linguistic knowledge contained in each model with respect to the others, by means of a set of calibrated experiments performed with the mentioned black-board system.

A parallel approach in statistical analysis of unrestricted corpora of human language

drs Jogchum Reitsma - Fryske Akademy

1 - Introduction

In this talk I will try to meet three goals, the last of which is the most important one. First of all I would like to say a few words about our institute, the Fryske Akademy (Frisian Academy); secondly I will introduce to you the project I am working on, that is the construction of a linguistic database of Frisian. The third and main goal is to sketch the plans we have for the automatic assignment of tags for word class and lemmatization, and the similarities and differences with other projects, aiming for the same goals.

2 - The Fryske Akademy

Being a small institute with only a short history in computer aided linguistic research, a short introduction of the Fryske Akademy will be useful, I think. Our institute was founded 1938 and has since occupied itself in a number of fields, notably, linguistics, regional history and social sciences, relating to the Dutch province of Friesland. It has its seat in Leeuwarden, the capital of Friesland, and employs some 60 researchers and administrative personnel. Amongst a large number of publications in the form of books, monographs and articles, a key publication is the Wurdboek fan de Fryske Taal, a scientific dictionary, of which up till now 7 of some 20 volumes have been published.

3 - The project

The computer aided linguistic research on the F.A. was until 1985 limited to some assistance in compiling the WFT, the above mentioned scientific dictionary. When, in 1985, the F.A. was given the opportunity of appointing some new researchers, the decision was taken to build a linguistic database of the Frisian language. A lexicologist, and an automation expert with a fair knowledge of syntax were invited to define and set up such an instrument, which was in the first place thought to be of further aid to the lexicographical staff, but had to be flexible enough to meet broader needs in e.g. the field of literature. I will describe the main choices we at the same time made in deciding which material to include, and in defining the technical system set-up. In a few words some of the results of our work so far (which will be described in a forthcoming paper) will be presented.

4 - Tagging the corpus

The next stage of the building of the LDB will incorporate the assignment of word class tags, and the lemmatization of the corpus. We are planning to adopt the basic method which has been used in tagging the well known LOB corpus.
In that project, undertaken at the University of Lancaster, a statistical approach was chosen. In quite a few publications (not only by the members of the Lancaster project, but also by linguists at the "Centrum voor taal en spraak" at the University of Nijmegen in The Netherlands, where this method has been used) this work has been described. The most thorough description is in Garside, Leech and Sampson, The computational analysis of English, London 1987. Chances are good that this method does not need further introduction, but if necessary I will give a short outline of the principles behind the system. Main keywords are statistical analysis, which has been proven adequate for tagging corpora of unrestricted human language.
After that I will desribe the differences between the Lancaster approach and ours. Some of these differences stem from variances in the history of our project (main difference is the fact that the Lancaster project could make use of an already hand-tagged corpus, the Brown corpus; we have nothing of this kind of material available); some others are caused by different choices in the technical set-up. These latter differences are two-fold: in the first place we put the result of the tagging of the corpus (i.e., the tags themselves) in a separate, commercially available database system. Right now we are using Oracle. It the second place we are aiming at a parallel approach. Parallellism in the LDB exists right now, but only in a rudimentary form; in the follow-up which I describe here we want a solid enhancement of parallellism.
In putting the results in a separate database we managed to get a flexible, yet fast environment for querying the corpus and updating it with new types of tagging. The choice for a relational database, apart from being a logical one in view of the market-acceptance of this system versus older (hierarchical or network) or newer (object-oriented, for example) database systems, makes it very easy to update the possibilities for a new type of information, gathered in a subproject. For example, at this moment we give positional information only in the form of a page/line number, and a fileposition. We think about enhancing positional data by offering the sentence/word number, thus allowing for querying for co-occurances of words at a given distance. In terms of the database layout this means only upgrading one table with two columns; a one-liner in the standard SQL database language. Also in querying the database the system is very flexible. Set up as a means to Keyword In Context (KWIC, pronounce as 'quick') information for the lexicographical staff, it is very easy to question the system for the words used by a certain author or group of authors in a certain period, or for words used in novels but not in newspapers. (Provided of course that these authors/periods/texttypes are represented in the corpus; we have spent considerable time and effort to reach a certain degree of representativeness; treatment of that aspect of our project is however beyond the scope of this paper; for this, an article decribing methods, successes and failures is in preparation.). An obvious advantage of a relational system is further that it has a wide impact in market. This gives a wide choice of modestly priced, well-behaving systems, running on a variety of hardware platforms and operating systems. One of the disadvantages of a relational system is its slowness, compared with other types of databases. However, the ever-increasing speed of hardware at lowering prices makes this a managable problem.
Since, as stated, we do not have an already handtagged corpus available, our tagging system has yet to be 'booted'. Altough human intervention seems indispensible, we would like to minimize it by creating - from an on-line available dictionary - a list of paradigmatized lemma's which, provided with its wordclass(es), is held against the word; all the human interventor has to do is select the correct wordclass, in case there is a choice

whatsoever. All statistical information, gathered during the tagging process, is of course kept in the database too. This means that we have an on-line update of these figures as the tagging goes on. This way we can get a faster and more accurate tagging: faster, because a word once tagged for a certain word class, needs not to be tagged again, provided the word class is the same; more accurate because the statistical information is kept accurate. At this moment parallellism is, as said, rather rudimentary. A stronger form of parallellism is chosen for the statistical arithmetic involved in the tagging. I will show the advantages of parallellism for this special problem with some examples, an after that explain the way parallellism is achieved. Keyword here is the well-known transputer. The word transputer is a contraction of transistor and computer, the device is a computer-on-a-chip, which needs almost only power supply and memory to operate. The at this moment strongest version of the chip is in computing power comparable with a 80386 with mathematical co-processor and caching memory and running on 25 MHz, but the special property of this chip is its ability to be connected to other transputers via four very fast on-chip links. I will show a number of connection schemes, and explain our choice in this respect. After that I will conclude my paper with discussing the way this transputersystem is incorporated in the rest of the system, and give some details about the way we will program the system. At this moment, it is not quite clear when we can start with the 'parallel' part of the project. The hardware (billed at some 65.000 DM for the aimed 25-transputersystem) has to be financed by the government, and, though so far we have got only positive responses, no final green light is given to order it. But chances are good, that by the time the congress is held, a first glimpse of our experiences can be given.

Unfortunately, there was no time for a native speaker of English to proofread this abstract, so please overlook the (possibly many) mistakes I have made.

A Statistical Approach for Phoneme to Grapheme Conversion

P. Rentzepopoulos, A. Tsopanoglou, G. Kokkinakis
Wire Communications Lab., University of Patras
265 00 Patras, GREECE
Tel.: +30 61 991722 TeleFax: +30 61 991855
e-mail: rentz@grpatvx1.bitnet

Abstract

The phoneme to grapheme conversion process and vice versa is common to man-machine communication systems (speech recognition, speech synthesis etc) and is basically the interface between the end user and the system.

The grapheme to phoneme conversion process (GTPC) has been implemented in the majority of the European languages mostly using rules. This method however has given poor results when used in the reverse process (i.e. phoneme to grapheme conversion - PTGC).

Statistical approaches have not been used widely in this area. The main reason is that the orthographic representation of speech in most European languages is supposed to be *following rules* something that is usually taught in spelling lessons at primary schools. In this paper the PTGC problem is approached from the statistical point of view using the theory of the *Hidden Markov Model*. The results produced from the experimentation with the proposed method are promising, since in the word level the system produced a score of 70% for unknown and 78% for known test corpus, while in the phoneme level the score was 94.6% and 95.5%, respectively.

I Introduction

The use of statistical approaches (MM, HMM, ML) has been generally accepted in the field of natural language processing (Language models based on 2-D Markov Models etc). These approaches are popular because they have the following advantages:

- Easy implementation,
- Good performance and robustness,
- Wide variety of applications,
- Fast response (since the training procedure is supposed to be done off-line).

In the field of PTGC these approaches have not been widely used. The most popular method is using algorithms based on phonological and heuristic rules. This procedure can be both

time consuming and inaccurate since the success of the overall system highly depends on the good knowledge of the user's mother tongue. The biggest disadvantage of this method is that the procedure for rule creation is usually by trial and error since, after a basic set of phonological rules is written, the system is tested and more rules are incorporated to deal with the exceptions of the previous set. In the phoneme to grapheme conversion process using rules, a set of some hundreds of rules may be produced from which only a few cover general aspects and most of them are dealing with special cases. Since the rules usually produce many candidates for each input phoneme, a geometrical increase of the amount of output suggestions with the size of the input word is observed. The biggest advantage of this method is that about 100% of the times the correct word will be included in the output list. Of course the function that describes the probability of existence of the correct word versus the number of words produced is monotonically decaying.

To overcome the problems of large output, a statistical approach to the problem of PTGC was used and specifically the *Hidden Markov Model (HMM)*. Considering the orthographic symbols as hidden states and the phonemic symbols as an observation sequence, the problem of Phoneme to Grapheme Conversion is to estimate a hidden state sequence that most likely produces the observation sequence, given the HMM model λ of the language.

II Theory

The Hidden Markov Model

The Hidden Markov Model is used to describe statistical phenomena that can be considered as sequences of hidden (unobservable) states which produce observable symbols. These phenomena are called hidden Markov Processes. The Hidden Markov process is described by a model λ which consists of three matrices A, B, π. The matrix A contains the transition probabilities of the hidden states, matrix B contains the probability of occurrence of an observation symbol given the hidden state and vector π contains the initial hidden state probabilities. If all the possible hidden states are N and all the observable events are M, then the dimensionality of matrix A is N by N, that of matrix B is N by M and π is a vector of N elements. The language is considered as a hidden Markov process with the orthographic form used as a hidden sequence of states (graphemes) and the phonemic form as the observation sequence.

B The algorithm

The problem that has to be addressed by the algorithm implemented is to calculate the best scored hidden state sequence, given the model λ and the observation sequence O. The main function of the algorithm is to find at any time (in the case of PTGC time is the position of a phoneme/grapheme in the word) the score of the best path in all possible hidden state sequences that end at the current state and recursively proceed from the beginning to the end of the word.

III Method

To implement an algorithm for the phoneme to grapheme conversion some decisions had to be taken about the hidden states, observation symbols and transition probabilities. These decisions are listed below.

- Every hidden state should produce exactly one observation symbol. To do so all the possible written forms of phonemes are coded as separate graphemic symbols (e.g. π and ππ are two different symbols even though they are both pronounced as /p/)
- The transition probability matrix (A) is biased to contain at least one occurrence of every transition so that it does not contain zero elements. This is done so that the algorithm does not discard a new transition but instead it assigns a bad score to it. This is perfectly allowed if the training corpus has a reasonably big size.

To measure the performance of the method two error criteria were defined:

a. Errors at word level
The system counts one error for every *word* which is not converted correctly to its graphemic form.

b. Errors at symbol level
The system counts one error for every *symbol* (unit grapheme) that does not match

with the input.

IV Results

The system was trained on various portions of three corpora. Finally it was trained with the total of these corpora. The texts that were used are:

- An office environment text of about 65000 words (OFFICE).
- Regulations and texts of Law of over 100000 words (LAW).
- A corpus of Greek newspaper articles consisting of more than 100000 words (NEWS).

Figure 1 summarises the actual sizes of the corpora and some attributes of the resulting HMM matrices.

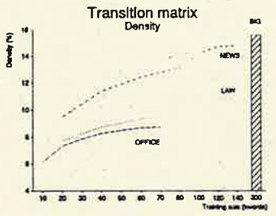


Figure 1

For the testing phase four texts were used. The first three were 10000 word portions of the forementioned texts and the fourth was a small corpus (ONEW) of about 4000 words from Greek newspapers not included in NEWS.

Figure 2 summarise the performance of the algorithms in various test conditions.

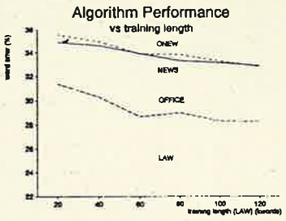


Figure 2

V Conclusions

The success rate of the algorithm is relatively high if the fact that the algorithm produces one output is considered. As it can be seen easily, the difference of the error rate at the word level versus the symbol level is justified from the fact that usually only one symbol is in error per word in error (average symbol error / word error ≈ 1.2) It is worth to note that the system does not use a dictionary for the conversion process, so it neither uses large amounts of computer memory nor has a long response time. Concluding, the advantages of the proposed method can be summarised as:

- The output of the algorithm is unique.
- The response time of the system is short enough since there is no lexical access.
- Many enhancements can be done depending on our needs.
- The algorithm is language independent.
- It does not have large memory requirements.
- The response time is linear to the word length while in the rule based systems is quadratic.

A self-organizing lexical system in hypertext

Burghard B. Rieger[0]
Constantin Thiopoulos
Department of Computational Linguistics
FB II: LDV/CL – University of Trier

1 Introduction

1.1 Knowledge-based Semantics

Our understanding of the bunch of complex intellectual activities subsumed under the notion of *cognition* is still very limited, particularly in how knowledge is acquired from texts and what processes are responsible for it. Recent achievements in wordsemantics, conceptual structuring, and knowledge representation within the intersection of cognitive psychology, artificial intelligence and computational linguistics have shown some agreement that cognition is (among others) responsible for, if not identifiable with, the processes according to which for a cognitive system previously unstructured surroundings may be tranformed to its perceived environment whose identifiable portions and their relatedness does not only constitute structures but also allow for their permanent revision according to the system's capabilities.
The common ground and widely accepted frame for modelling the semantics of natural language is to be found in the dualism of the rationalistic tradition of thought as exemplified in its notions of some independent (objective) reality and the (subjective) conception of it. According to this *realistic* view, the meaning of a language term (i.e. text, sentence, phrase, word, syllable) is conceived as something being related somehow to (and partly derivable from) certain other entities, called signs, a term is composed of. As a sign and its meaning is to be related by some function, called interpretation, language *terms*, composed of *signs*, and related *meanings* are understood to form some structures of entities which appear to be at the same time part of the (objective) reality and its (subjective) interpretation of it. In order to let signs and their meanings be identified as part of language terms whose interpretations may then be derived, some knowledge of these structures has to be presupposed and accessible for any symbolic information processing. Accordingly, *understanding* of language expressions can basically be identified with a of matching some input strings with supposedly predefined configurations of word meaning and/or world structure whose representations have to be available to the (natural or artificial) understanding system's particular (though limited) *knowledge*. The so-called *cognitive paradigm* of advanced procedural linguistics can easily be traced back to stem from this fundamental duality, according to which natural language understanding can be modelled as the *knowledge-based* processing of information.

[0]partly by support of *The German Marshall Fund of the United States*

Subscribing to this notion of understanding, however, tends to be tantamount to accepting certain unwarranted presuppositions of theoretical linguistics (and particularly some of its model-theoretical semantics) which have been exemplified elsewhere[1] by way of the formal and representational tools developed and used so far in cognitive psychology (*CP*), artificial intelligence (*AI*), and computational linguistics (*CL*). In accordance with these tools, *word meaning* and/or *world knowledge* is uniformly represented as a (more or less complex) labelled graph with the (tacid) understanding that associating its vertices and edges with symbols from some established system of sign-entity-relationship (like e.g. that of natural language) will render such graph-theoretical configurations a model of structures or properties which are believed to be those of either the sign-system that provided the graphs' labels or the system of entities that was to be depicted. Obviously, these representational formats are not meant to model the *emergence* of structures and the *processes* that constitute such structures as part of word meaning and/or world, but instead are merely making use of them[2].

1.2 Cognitive Semiotics

It has long been overlooked that relating arc-and-node structures with sign-and-term labels in symbolic knowledge representation formats is but another illustration of the traditional *mind-matter*-duality presupposing a realm of *meanings* very much like the structures of the *real world*. This duality does neither allow to explain where the structures nor where the labels come from. Their emergence, therefore, never occurred to be in need of some explanatory modelling because the existence of *objects, signs* and *meanings* seemed to be out of all scrutiny and hence was accepted unquestioned. Under this presupposition, fundamental *semiotic* questions of *semantics*—simply did not come up, they have hardly been asked yet[3], and are still far from being solved.
In following a *semiotic paradigm* this inadequacy can be overcome, hopefully allowing to avoid (if not to solve) a number of spin-off problems, which originate in the traditional distinction and/or the methodological separation of the meaning of a language's term from the way it is employed in discourse. It appears that failing to mediate between these two sides of natural language semantics, phenomena like *creativity, dynamism, efficiency, vagueness,* and *variability* of meaning—to name only the most salient—have fallen in between, stayed (or be kept) out of the focus of interest, or were being overlooked altogether, sofar. Moreover, there is some chance to bridge the gap between the formal theories of language description (*competence*) and the empirical analysis of language usage (*performance*) that is increasingly felt to be responsible for some unwarranted abstractions of fundamental properties of natural languages.
Approaching the problem from a *cognitive* point-of-view, identification and interpretation of external structures has to be conceived as some form of *information processing* which (natural/artificial) systems—due to their own structuredness—are (or ought to be) able to perform. These processes or the structures underlying them, however, ought to be derivable from—rather than presupposed to—procedural models of meaning[4]. Based upon a phenomenological reinterpretation of the analytical concept of *situation* as expressed by BARWISE/PERRY (1983) and the synthetical notion of *language game* as advanced by the late WITTGENSTEIN (1958), the combination of both lends itself easily to operational extensions in empirical analysis and procedural simulation of associative meaning constitution which may grasp essential parts of what PEIRCE named *semiosis*[5].
Modelling the meaning of an expression along reference-theoretical lines had to presuppose the structured sets of entities to serve as range of the denotational function which provided the expression's interpretation. However, it appears feasible to have this very range be constituted as a result of exactly those cognitive functions by way of which understanding is produced. It will have to be modelled as a dynamic generation which reconstructs the possible structural connections of an expression towards cognitive systems (that may both intend/produce and realize/understand it) and in respect to their *situational* settings, being specified by the expressions' pragmatics.
In phenomenological terms, the set of structural constraints defines any cognitive (natural or artificial) system's possible range in constituting its schemata whose instantiations will determine the system's actual interpretations of what it perceives. As such, these cannot be characterized as a domain of objective entities, external to and standing in contrast with a system's internal, subjective domain; instead, the links between these two domains are to be thought of as *ontologically fundamental*[6] or pre-theoretical. They constitute—from a *semiotic* point-of-view—a system's primary means of access to and interpretation of what may be called its "world" as the system's particular apprehension of its environment. Being fundamental to any cognitive activity, this basal identification appears to provide the grounding framework which underlies the duality of categorial-type rationalistic mind-world or subject-object separation.
From a systems-theoretical point-of-view, this is tantamount to a shift from linear to non-linear systems in modelling cognitive and semiotic behaviour. The simplest way to distinguish these approaches is by identifying the behaviour of *linear systems* as being equal to the sum of the behaviour of its parts, whereas the behaviour of *non-linear systems* is more than than the sum of its parts. FREGEs principle of *compositionality* as well as CHOMSKEYs hypotheses of indepedance of syntax are concepts in point of the *linear*-systems'-view: by studying the parts of a system in isolation first, will then allow for a full understanding of the complete system by composition. This collides with the *non-linear*-systems'-view according to which the primary interest is not in the behaviour of parts as properties of a system but rather in the behaviour of the *interaction* between parts of a system. Such interaction-based properties necessarily disappear when the parts are studied in isolation, as can be witnessed in referencial and model-theoretic semantics where phenomena like *vagueness, contextual variability* and *creative dynamism* cannot be dealt with, or in competence theoretical syntax where grades of *grammaticality, adaptive change* and *discourse adequacy* cannot be addressed.
The *self-organizing* property of the non-linear system introduced here has formally been derived elsewhere[7] from mathematical *topos theory*[8] and *category theory*[9]. This implementation of the system and its organisation as a dynamic *hypertext* structure is to simulate the emergence of lexical meanings by way of word co-occurrence constrains of—as yet—rather coarse syntagmatic/paradigmatic regularities in natural language texts.

[1]Rieger 1991
[2]For illustrative examples and a detailed discussion see Rieger 1989, pp.103–132.
[3]see however Rieger (1977)
[4]It has been argued elsewhere (Rieger 1990, 1991) that *meaning* need not be introduced as a presupposition of *semantics* but may instead be derived as a result of semiotic modelling.
[5]"By *semiosis* I mean [...] an action, or influence, which is, or involves, a coöperation of *three* subjects, such as sign, its object, and its interpretant, this tri-relative influence not being in any way resolvable into actions between pairs." (Peirce 1906, p.282)
[6]Heidegger (1927)
[7]Thiopoulos 1992 forthcoming
[8]Goldblatt 1979
[9]Bell 1981; Lambek/Scott 1986
[10]For a complete description of the theoretical framework see (Thiopoulos,1990).

2 The formalism

2.1 The self-organizing mechanism

A numerical measure expressing the dependency between two lexems can be calculated by taking the number of common contexts to be a representation of their mutual use. Thus for $\mathcal{O}(a)$ set of contexts of a, i.e. texts, where an instantiation of a appears, we define:

Definition 2.1 $conf(a,b) = \frac{|\mathcal{O}(a) \cap \mathcal{O}(b)|}{|\mathcal{O}(b)|}$

For L set of the considered lexems, the actual state of the structure is given by a matrix $CONF = (conf(a,b))_{a,b \in L}$. The self-organizing modification can be obtained by recomputing $conf$ after each new context. There are three cases:

1. a and b are both in the new text. Then: $conf(a,b)_{new} = \frac{|\mathcal{O}(a) \cap \mathcal{O}(b)|+1}{|\mathcal{O}(b)|+1} = \frac{\frac{|\mathcal{O}(a) \cap \mathcal{O}(b)|+1}{|\mathcal{O}(b)|}}{\frac{|\mathcal{O}(b)|+1}{|\mathcal{O}(b)|}} = \frac{conf_{old} + \frac{1}{|\mathcal{O}(b)|}}{1+\frac{1}{|\mathcal{O}(b)|}}$. In this case $conf(a,b)_{new} \geq conf(a,b)_{old}$, i.e the intensity of the connection between a and b increases.
2. Only b is in the new text. Then: $conf(a,b)_{new} = \frac{conf_{old}}{1+\frac{1}{|\mathcal{O}(b)|}}$. In this case $conf(a,b)_{new} \leq conf(a,b)_{old}$, i.e the intensity of the connection between a and b decreases.
3. Only a is in the new text. Then: $conf(a,b)_{new} = conf(a,b)_{old}$.

2.2 Categories

In order to capture structural features of the actual state of the system the $CONF$ matrix is transformed to a category[10]. A category is a directed graph with some additional features.

Definition 2.2 *A category A consists of:*

1. *a class of objects OBJ(A)*
2. *a class of morphisms MORPH(A)*
3. *two operations dom,cod:MORPH(A)→OBJ(A), with f:a → b* iff *dom(f)=a and cod(f)=b*
4. *an operation comp : MORPH(A)×MORPH(A) → MORPH(A) with comp(f,g)=f∘g such that* $f \circ (g \circ h) = (f \circ g) \circ h$
5. *an operation id:OBJ(A) → MORPH(A) with* $id(a) = id_a : a \to a$, *where* id_a *is the identity function on a .*

The nodes of the graph are the objects and the links the morphisms. For $f : a \to b$, a is the domain of f and b the codomain. *comp* is the (associative) composition of morphisms and *id* maps each object to the corresponding identity morphism.

Definition 2.3 $A(a,b) = \{f \mid f \in MORPH(A) \wedge f : a \to b\}$

Definition 2.4 *B is a subcategory of* $A(B \subseteq A)$ *iff*

1. $OBJ(B) \subseteq OBJ(A)$
2. $\forall a,b \in OBJ(B) B(a,b) \subseteq A(a,b)$.

Definition 2.5 *B is a full subcategory of A iff*

1. $B \subseteq A$
2. $\forall a,b \in OBJ(B) B(a,b) = A(a,b)$.

A full subcategory is thus a subcategory that contains all the morphisms of the original category between its objects, i.e. it is a function closed under functional application.
A special class of categories is the class of cartesian closed categories. They are characterized by the fact that some structural operation are defined of them. Here we consider two of them:

Definition 2.6 *The product from* $a,b \in OBJ(A)$ *is* $a \times b \in OBJ(A)$ *together with* $pr^1 : a \times b \to a, pr^2 : a \times b \to b \in MORPH(A)$, *so that* $\forall c \in OBJ(A) \exists! < f,g >: c \to a \times b \in MORPH(A)$ *with* $pr^1 \circ < f,g >= f \wedge pr^2 \circ < f,g >= g$.

Definition 2.7 *The coproduct from* $a,b \in OBJ(A)$ *is* $a+b \in OBJ(A)$ *together with* $\kappa^1 : a \to a+b, \kappa^2 : b \to a+b \in MORPH(A)$, *so that* $\forall c \in OBJ(A) \exists! (f,g) : a+b \to c \in MORPH(A)$ *with* $(f,g) \circ \kappa^1 = f \wedge (f,g) \circ \kappa^2 = g$.

The transformation of the matrix $CONF$ to a category $C(CONF)$ is defined by:

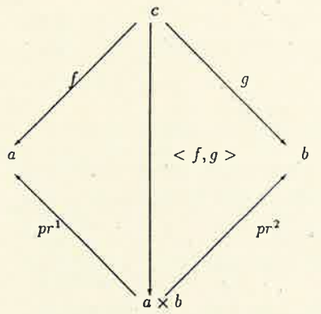


Figure 1: Product

- $OBJ(C(CONF)) = L$
- $conf(a,b) \geq conf(b,a) \Rightarrow f : a \to b \in MORPH(C(CONF))$
- $conf(a,b) < conf(b,a) \Rightarrow f : b \to a \in MORPH(C(CONF))$

The weighting of the morphisms is thus given as a partial function:

$$conf(f) = conf(a,b) \text{ iff } dom(f) = a \wedge cod(f) = b \wedge conf(a,b) \geq conf(b,a).$$

that can be extended to morphism combination as follows:

for the composition: $conf(f \circ g) = conf(f) conf(g)$

for the product: $conf(< f,g >) = minimum(conf(f), conf(g))$.

for the coproduct: $conf((f,g)) = minimum(conf(f), conf(g))$.

The meaning of a lexem a, as a structural description of how a is interlinked in the network of lexems, according to a numerical boundary GLB that determines the depth of the activation, is given by:

Definition 2.8 $a^{*_{GLB}} = \{(b, conf(f)) \mid \exists f \in MORPH(ST) f : a \to b \wedge conf(f) \geq GLB\}$

The meaning of two or more lexems can be represented as a full subcategory generated by the product and coproduct constructions.

Definition 2.9 $prod_{GLB}(a,b) = \{(C, conf(f)) \mid \exists f : C \to a \times b \wedge conf(f) \geq GLB\}$
$coprod_{GLB}(a,b = \{(C, conf(f)) \mid \exists f : a+b \to C \wedge conf(f) \geq GLB\}$

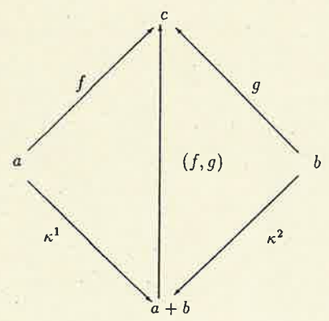


Figure 2: Coproduct

Definition 2.10 *The situation generated out of* a,b *in a cartesian closed category C is the full subcategory* $SIT(a,b)$ *with*

$$OBJ(SIT(a,b)) = \{c \in OBJ(C) \mid c \text{ collected by } prod(a,b) \text{ and } coprod(a,b) \text{ for a specific } GLB\}.$$

A situation is thus a substructure of the original category that is closed under functional application and since it is a category again it is possible to apply the same mechanisms as in the original category. The meaning of a lexem *a relative* to a situation, $SIT(b,c)$ is a^* in $SIT(b,c)$.

3 The Implementation

Hypertext seems to be the most suitable tool for capturing the dynamic nature of the formalism. The category $C(CONF)$ can, as a directed graph, be mapped into a hypertext structure[11], in the following way:

-Each object of $C(CONF)$ is implemented as a card named after this object that contains a text field with the name (see figure 3).

- Each morphism $f : a \to b$ is implemented as a button on the card named a that leads, when activated, to the card named b. The name of the button is formed by concatenating b with $conf(f)$ (see figure 3). The user can navigate through the network by clicking on the buttons.

- The structural operations defined on $C(CONF)$ can be implemented as *browsers* and the determined portions of the network can be accessed via *hyperviews*.

Besides the cards that correspond to the objects of $C(CONF)$ there is also a control card, from where the user controls the implemented navigation mechanisms (see figure 4). These mechanisms can be activated by clicking on the buttons of the control card:

1. *Read Text* calls a C program that reads the actual text and recomputes the $CONF$ matrix.
2. *Topos* generates from the $CONF$ matrix the corresponding hypertext file.
3. *View* produces a global view of the file.
4. *Interprct* generates $*_{GLB}$ for the lexem that is given in the left text field at the top of the card (industrie), where the depth of the activation (GLB) can be specified by the user as entry in a text field (here is 0.5). *Interprct* activates a browser that avoids cycles by keeping a list of visited cards. The collected lexems are listed, together with the corresponding weight and card number, in the left scrolling field in the center of the card.
5. *Situation* activates the *prod* and *coprod* mechanisms for the lexem contained in the left text field at the top of the card (industrie) and the lexem (or lexems) contained in the scrolling field at the top of the card (technik). The right srolling field in the center of the card is tereby used to keep track of the lexems acessed by the different interpretations.
6. *Full* generates the corresponding full subcategory (here sit(industrie,technik), i.e. determines all the morphisms between the collected lexems.
7. *Go to situation* leads to the control card of a hypertext file that corresponds to the actual situation, where by using the *Topos* button the full subcategory is mapped - using the same mapping operation as for $C(CONF)$ - into this file.

The process of restricting the category $C(CONF)$ can be used recursively and reflects the focus of interest of the user of the system. In this example (that, since it corresponds to the first stages of the system has a rough structure) the view of the category $C(CONF)$ is given in figure 5 and the view of the full subcategory $sit(industrie, technik)$ is given in figure 6. The text field in figure 6 contains the lexems that led to this view and for successive determinations of situations, i.e. situations of situations of ...,it contains the history of the restrictions. By using the *Interpret* button of the hypertext file that represents the actual situation the user can determine the meaning of a lexem *relative* to this situation. This reflects the philosophy of the user/system integration immanent in the hypertext approach and leads towards a solution of the problem of "getting lost" that prevents hypertext from exploiting its full power as a flexible and user friendly tool for the sophisticated use of electronically stored information.

[11]The implementation is made in Hypercard

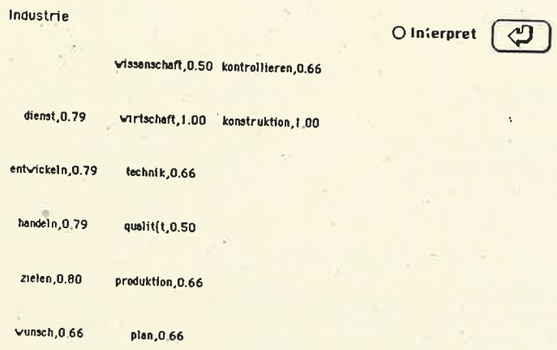


Figure 3: A card containing a text field with the name of the corresponding lexem and buttons with the names of the lexems that are codomains of morphisms starting from this lexem.

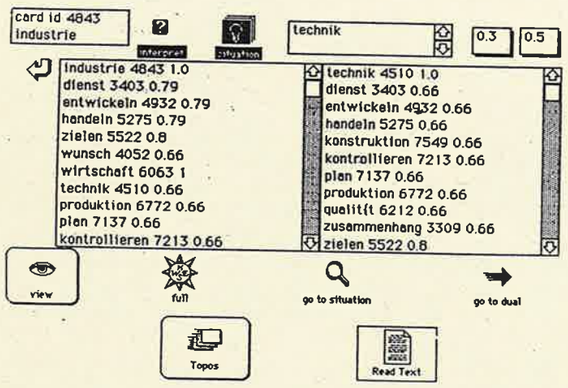


Figure 4: The control card

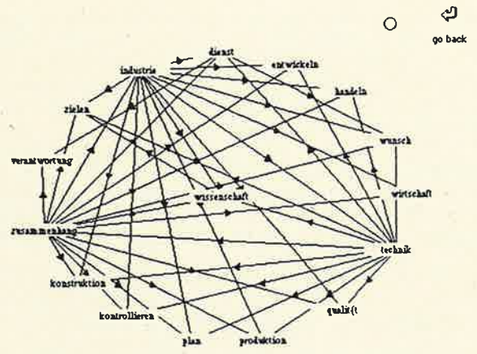


Figure 5: The view of the hypertext file representing $C(CONF)$

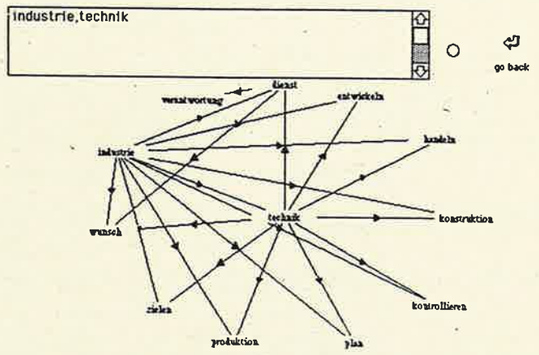


Figure 6: The view of the hypertext file representing $sit(industrie, technik)$

References

Barrett, E. (Ed.)(1988): *Text, ConText, and HyperText.* Cambridge, MA (MIT)

Barwise, J./ Perry, J.(1983): Situations and Attitudes. Cambridge, MA (MIT)

Bell, J. L. (1981): "Category theory and the foundation of mathematics" *British Journal of the Philosophy of Science,* 32,1981:349–358

Conklin,J. (1987): "Hypertext: an introduction and survey" *Computer*, Vol.20, No.9.

Frisse, M. (1987): "From text to hypertext" *Byte* October, 1987.

Goldblatt, R. (1984): Topoi. The Categorial Analysis of Logic. (Studies in Logic and the Foundations of Mathematics 98), Amsterdam (North Holland)

Heidegger, M. (1927): Sein und Zeit. Tübingen (M.Niemeyer)

Lambek, J./ Scott P. J. (1986): Introduction to higher order categorical logic. Cambridge (Cambridge University Press)

Langdau, C. G. Artificial Life. In Langdau, C. G. (ed.). *Artificial Life. Proceedings of an interdisciplinary workshop 1987.* Los Alamos, Redwood City/Menlo Park, CA 1989, pp. 1-48.

Maturana, H./ Varela, F. (1980): Autopoiesis and Cognition. The Realization of the Living. Dordrecht (Reidel)

Peirce, C.S. (1906): "Pragmatics in Retrospect: a last formulation" (CP 5.11 – 5.13), in: The Philosophical Writings of Peirce. Ed. by J. Buchler, NewYork (Dover), pp.269–289

Rieger, B. (1977): "Bedeutungskonstitution. Einige Bemerkungen zur semiotischen Problematik eines linguistischen Problems" *Zeitschrift für Literaturwissenschaft und Linguistik* 27/28, pp.55–68

Rieger, B. (1985b): "On Generating Semantic Dispositions in a Given Subject Domain" in: Agrawal, J.C./ Zunde, P. (Eds.): Empirical Foundation of Information and Software Science. NewYork/ London (Plenum Press), pp.273–291

Rieger, B. (1989): Unscharfe Semantik. Die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten. Frankfurt/ Bern/ NewYork (P. Lang)

Rieger, B. (1990): "Situations and Dispositions. Some formal and empirical tools for semantic analysis" in: Bahner, W. / Schildt, J./ Viehweger, D. (Ed.): Proceedings of the XIV.International Congress of Linguists (CIPL), Vol. II, Berlin (Akademie Verlag), pp.1233–1235

Rieger, B. (1991): "On Distributed Representation in Word Semantics" (ICSI-Report TR-91-012) International Computer Science Institute, Berkeley, CA

Rieger, B.B./ Thiopoulos, C. (1989): Situations, Topoi, and Dispositions. On the phenomenological modelling of meaning. in: Retti, J./ Leidlmair, K. (Eds.): 5th Austrian Artificial Intelligence Conference. (ÖGAI 89) Innsbruck; (KI–Informatik-""Fachberichte Bd.208) Berlin/ Heidelberg/ NewYork (Springer), pp.365–375

Thiopoulos, C. (1991): Meaning metamorphosis in the semiotic topos. *Theoretical Linguistics*16, 2/3, 255-274, 1990.

Winograd,T./ Flores, F. (1986): Understanding Computers and Cognition: A New Foundation for Design. Norwood, NJ (Ablex)

Wittgenstein, L. (1958): The Blue and Brown Books. Ed. by R. Rhees, Oxford (Blackwell)

Pauli Saukkonen

MAIN TRENDS AND RESULTS OF QUANTITATIVE LINGUISTICS IN FINLAND

Linguistic research in Finland began to utilize quantitative methods more widely in the 1960s owing to the influence and expansion of automatic data processing on the one hand and social and behavioural sciences on the other. In the contemporary monographs, quantification of phenomena is rather the rule than an exception. It is used in phonetic-phonological, morphological, lexical, syntactic and textual studies. I shall concentrate on those in which quantitative methods receive the main emphasis.

Quantitative linguistics has been and still is mostly statistical by nature. Statistical data are very essential in sociolinguistic and stylistic research. There are many projects in this field. It has been proved that all sorts of phenomena from sounds to grammatical categories may have different frequency distribution in different texts and discourses. All linguistic variables are hence principal and potential stylistic markers. This was not a general opinion earlier.

Factor analysis has been used to search and discriminate the interpretations why linguistic categories behave differently. The results are different from the factors of Douglas Biber (Variation across speech and writing, Cambridge 1988).

Synergetic interplay between the length and the frequency of words, clauses and sentences prevails universally also in Finnish. Some questions still remain, why certain exceptions in certain text types exist. One interesting finding is the "Wannmühlegesetz": in clauses of different lengths, the words almost invariably grow longer towards the end. This has some synergetic explanations.

Many different materials have been used: texts and morphological and lexical collections from areal and social dialects for phonological, morphological, lexical and syntactic studies, texts from children and schoolchildren for language acquisition studies, texts from spoken and written standard language, from members of the Parliament, radio, TV, newspapers, fiction and non-fiction for stylistics, text type and grammatical studies, frequency vocabularies and probabilistic and stochastic modelling, the vocabulary of current Finnish for the international project Language Synergetics.

Die Behandlung von Reparaturen in konnektionistischen Sprachproduktionsmodellen
– Extended Abstract –

Ulrich Schade
Uwe Laubenstein
Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld

1. Reparaturen

Bei der Produktion gesprochener Sprache unterlaufen dem Sprecher gelegendlich tatsächliche oder vermeintliche Fehler, die er, sofern sie von ihm bemerkt werden, zumeist zu reparieren sucht. Ein typisches Beispiel einer Reparatur findet sich in (1).

(1) „und vorne drauf liegt ein grünes eh n blaues Rechteck"
(Forschergruppe Kohärenz 1987)

Reparaturen sind erst in neuerer Zeit zu einem allgemein anerkannten Gegenstand linguistischer Untersuchungen geworden. Einen ersten Ansatz für ihre systematische Analyse stellen die Arbeiten von Schegloff, Jefferson und Sacks (1977) und Schegloff (1979) dar. In diesen Arbeiten wird der Begriff „Reparatur" („repair") eingeführt. Des weiteren wird versucht, den Gegenstandsbereich einzugrenzen und zu gliedern. Ein wichtiges Ergebnis dieser Arbeiten war die Erkenntnis, daß Reparaturen in ihrer syntaktischen Struktur nicht willkürlich sondern offenbar nach bestimmten festen Regeln durchgeführt werden.

Levelt (1983) vertiefte das Wissen über Reparaturen, indem er für sogenannte Selbstreparaturen (bei denen lediglich der Sprecher nicht jedoch sein Gesprächspartner an der Einleitung und Durchführung der Reparatur beteiligt ist) eine verfeinerte Typologie vorschlug und Reparaturen syntaktisch in einen Zusammenhang mit Koordinationskonstruktionen stellte.

2. Levelts Koordinationsthese für Reparaturen

Aufgrund der Beobachtung, daß Reparaturen in ihrer Oberflächenrealisierung zumeist Koordinationskonstruktionen entsprechen, formuliert Levelt (1983) folgende Wohlgeformtheitsregel für Reparaturen:

> „A repair $<\alpha\gamma>$ is well-formed if and only if there is a string β such that the string $<\alpha\beta$ and* $\gamma>$ is well-formed, where β is a completion of the constituent directly dominating the last element of α. (* *and* to be deleted if γ's first element is itself a sentence connective)." (S. 78)

Diese Regel ist für die meisten Reparaturen sicherlich korrekt; sie gilt aber – wie Levelt selbst betont (1983, S. 81) – nicht für Reparaturen, die sich auf die syntaktische Form der Äußerung beziehen. Folgende Reparaturen werden also nicht durch die Regel abgedeckt:

(2) „En zwart ... van zwart naar rechts naar rood" (Levelt 1983)
(3) „weil ein bißchen soll auch die Vordergrund – die Musik im Vordergrund stehen" (Berg 1987)
(4) „Und dann nimmst Du dem ähh den roten Klotz"

Levelts Wohlgeformtheitsregel ist ein Richtmaß dafür, ob eine Reparatur anhand ihrer Oberflächenkonstruktion als wohlgeformt klassifiziert werden kann oder nicht. Es stellt sich aber die Frage, ob die Regel auch auf einem kognitiven Prinzip beruht; ob sie also ein Verfahren angibt, nach dem von natürlichen Sprechern Reparaturen produziert werden, wie dies in Levelt (1989) anklingt:

> „Syntactically speaking, an utterance and its repair constitute a kind of *coordination* (Levelt 1983; DeSmedt and Kempen 1987), and the syntactic rules of coordination have to be followed. This state of affairs can be captured in a *Well-Formedness Rule* for repairs." (Levelt 1989, S. 486)

Bevor wir versuchen wollen, diese Frage zu beantworten, soll aufgezeigt werden, wie die Fähigkeit zur Selbstreparatur in ein konnektionistisches Sprachproduktionsmodell integriert werden kann, da die Betrachtung des Modells Hinweise zur Beantwortung dieser Frage liefert.

3. Ein konnektionistisches Sprachproduktionsmodell

Für die kognitive Fähigkeit der Sprachproduktion von Erwachsenen existieren in der Literatur konnektionistische Modelle in der Tradition von Gary Dell (Dell & Reich 1980; Dell 1985, 1986, 1988; Stemberger 1985a, 1985b, 1990; Berg 1986, 1988; MacKay 1987; Schade 1988, 1990a). Die verschiedenen Modelle unterscheiden sich in der Repräsentation syntaktischer Regeln und der damit verbundenen Realisation der Sequentialisierung linguistischer Einheiten wie Wörter, Silben und Phoneme während des Produktionsprozesses. Alle diese Modelle sind lokale konnektionistische Modelle (nach der Terminologie von Rumelhart und McClelland 1986), da sie jede der genannten linguistischen Einheiten, also jedes Wort, jede Silbe, jedes Phonem aber auch jedes phonologische Merkmal, durch einen Knoten eines konnektionistischen Netzwerks repräsentieren.

Die Arbeiten zur Entwicklung eines konnektionistischen Produktionsmodells wurden von der DFG im Rahmen der DFG-Forschergruppe „Kohärenz" an der Universität Bielefeld gefördert.

Grundsätzlich bestehen konnektionistische Modelle aus einem Netzwerk miteinander verbundener Knoten, die über einen Aktivierungswert verfügen. Übersteigt der Aktivierungwert eines Knotens seinen Schwellwert, so kann der Knoten an seine Nachbarn Aktivierung weiterleiten. Die Menge der Aktivierung, die ein Knoten weiterleitet, hängt dabei von zweierlei Faktoren ab: Sie ist zum einen proportional zu der Aktivierung des sendenden Knotens, und sie ist zum anderen proportional zu der Stärke der Verbindung zwischen dem sendenden Knoten und dem Empfänger. Aufgrund der eingehenden Aktivierung verändern sich die Aktivierungswerte der Knoten.

Die Knoten der konnektionistischen Modelle für die Sprachproduktion sind in Ebenen angeordnet. Es gibt beispielsweise eine Ebene mit Wortknoten, eine Ebene mit Silbenknoten, eine Ebene mit Phonemknoten und mehrere Ebenen mit Knoten für phonologische Merkmale, etwa Ebenen für Stimmhaftigkeit, für Artikulationsorte und für Artikulationsarten. In dem hier zugrunde gelegten Modell sind alle Knoten einer Ebene paarweise durch inhibitorische Leitungen verbunden. Inhibitorische Leitungen sind Verbindungen zwischen Knoten, die zu einer Verkleinerung des Aktivierungswertes beim empfangenden Knoten führen, ihn also hemmen. Inhibitorische Verbindungen ergeben sich aus paradigmatischen Relationen und beschreiben Wahlmöglichkeiten für bestimmte strukturelle Positionen: Zu einer bestimmten Zeit kann immer nur ein Wort, nur eine Silbe und nur ein Phonem geäußert werden.

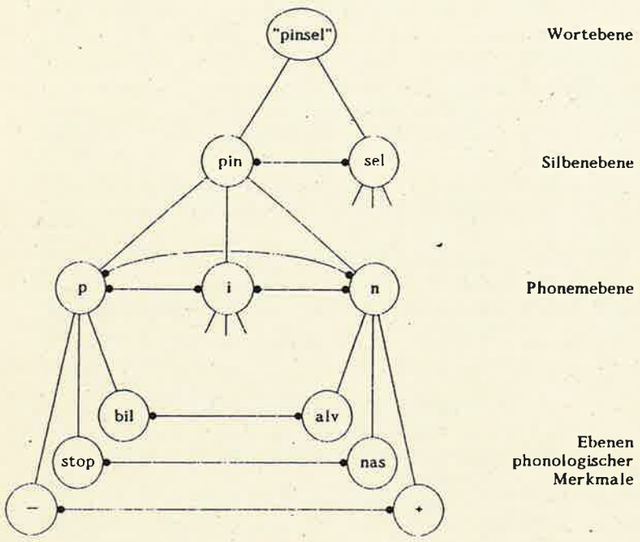


Abbildung 1: Ausschnitt aus einem konnektionistischen Netzwerk zur Sprachproduktion

Anders als bei den inhibitorischen Leitungen, die von Dell und MacKay nicht vorgesehen sind, stimmen die Modelle bezüglich der exzitatorischen Leitungen (Verbindungen, über die Aktivierung weitergeben wird, die zur Vergrößerung des Aktivierungswertes des empfangenden Knotens führt) überein. Exzitatorische Leitungen ergeben sich aus syntagmatischen Relationen zwischen den linguistischen Einheiten und verbinden stets Knoten benachbarter Ebenen: beispielsweise ist ein Silbenknoten mit denjenigen Knoten der Phonemebene verbunden, die die Phoneme der entsprechenden Silbe repräsentieren (vgl. Abbildung 1). Alle exzitatorischen und inhibitorischen Verbindungen sind symmetrisch, so daß bei einer Produktion Feedbackeffekte auftreten (vgl. Dell 1985).

Wird mit der Hilfe eines konnektionistischen Netzwerkes der beschriebenen Art Sprachproduktion modelliert, so wird derjenige Knoten auf einen hohen Aktivierungswert gesetzt, der die zu äußernde linguistische Einheit repräsentiert. Soll beispielsweise ein Wort geäußert werden, wäre dies der entsprechende Knoten der Wortebene.

Die empirischen Erkenntnisse, die über den kognitiven Prozeß der Sprachverarbeitung vorliegen, erfordern die sequentielle Realisierung der Phoneme eines Wortes (vgl. z.B. Meyer 1988). In einem konnektionistischen Modell wird daher in bestimmten kurzen zeitlichen Abständen ermittelt, welches jeweils der höchst aktivierte Phonemknoten ist bzw. welches die höchst aktivierten Merkmalknoten der betroffenen Ebenen von phonologischen Merkmalen sind. Diese Knoten legen das Produktionsresultat für den betreffenden Zeitpunkt fest und werden anschließend durch eine Selbstinhibition auf den Aktivierungswert „0" zurückgesetzt.

Für die Sequentialisierung der linguistischen Einheiten ist in dem zugrunde gelegten Modell ein spezielles Teilnetz zuständig, in dem syntaktische Regeln durch sogenannte „Kontrollknotenketten" repräsentiert werden (vgl. Schade 1988, Schade 1990a, Eikmeyer & Schade 1991). Eine typische Kontrollknotenkette, mit der etwa die Produktion einer Nominalphrase realisiert werden könnte, besteht aus den Knoten `DET`, `ADJ` und `N`. Sofern eine solche Kette aktiviert wird, hat immer genau einer der genannten Knoten den Aktivierungswert „1" und die anderen den Wert „0". Während der Produktion kann derjenige Knoten, der aktivert ist, Wortknoten seiner syntaktischen Kategorie unterstützen. Nachdem ein entsprechendes Wort produziert wurde, gibt der aktivierte Knoten der Kette seine Aktivierung an seinen Nachfolger weiter und hemmt sich selbst. Durch dieses Verfahren kann die Sequentialisierung realisiert werden.

Sofern das beschriebene Modell ‚Indizien' für eine fehlerhafte Produktion ‚erkennt' (zur Monitorkomponente konnektionistischer Produktionsmodelle vgl. Berg 1986, McKay 1987, Schade 1990b), leitet es einen Reparaturversuch ein. Dabei wird die Aktivierung in den Kontrollknotenketten anhand eines Algorithmus „zurückgesetzt". Anschließend wird die Produktion lediglich wieder aufgenommen. Dieses einfache Verfahren bedingt eine normalerweise erfolgreiche Reparatur. Die entsprechenden Repraraturresultate entsprechen dabei denjenigen eines natürlichen Sprechers.

4. Die Folgerung aus dem Modell für Levelts Koordinationsthese

Das im Modell realisierte Verfahren zur Durchführung von Reparaturen greift einerseits nicht auf die Prinzipien (diejenigen Kontrollknotenketten) zurück, mit denen Koordinationen realisiert werden, wobei die Produktionsresultate normalerweise trotzdem der von Levelt formulierten Wohlgeformtheitsregel genügen. Andererseits lassen sich mit dem Verfahren auch Fehler syntaktischer Art reparieren.

Will man Reparaturen nicht in zwei Klassen teilen, wobei die Mitglieder der einen Klasse kognitiv so realisiert werden wie Reparaturen, die Mitglieder der anderen Klasse jedoch nicht, so kann man in Levelts Wohlgeformtheitsregel zwar eine Regel sehen, die für die Klassifikation von Reparaturen anhand von Oberflächenmerkmalen nützlich ist, die aber nicht als kognitives Prinzip verstanden werden sollte. Das vorgestellte Modell zeigt, daß es möglich ist, ein Verfahren zu formulieren, nach dem alle Reparaturtypen realisiert werden können und das trotzdem mit den empirischen Daten in Einklang steht.

Literaturverzeichnis

Berg, T. (1986). The Problems of Language Control: Editing, Monitoring, and Feedback. *Psychological Research*, 48, 133–144.

Berg, T. (1987). The Case against Accommodation: Evidence from German Speech Error Data. *Journal of Memory and Language*, 26, 277–299.

Berg, T. (1988). *Die Abbildung des Sprachproduktionsprozesses in einem Aktivierungsflußmodell.* Tübingen: Niemeyer.

Dell, G.S. (1985). Positive Feedback in Hierarchical Connectionist Models: Applications to Language Production. *Cognitive Science*, 9, 3–23.

Dell, G.S. (1986). A Spreading-Activation Theory of Retrieval in Sentence Production. *Psychological Review*, 93, 283–321.

Dell, G.S. (1988). The Retrieval of Phonological Forms in Production: Tests of Prediction from a Connectionist Model. *Journal of Memory and Language*, 27, 124–142.

Dell, G.S. & Reich, P.A. (1980). Toward a Unified Model of Slips of the Tongue. In: V.A. Fromkin (Hrsg.), *Errors in Linguistic Performance*, 273–286. New York, NY: Academic Press.

DeSmedt, K. & Kempen, G. (1987). Incremental Sentence Production, Self-Correction, and Coordination. In: G. Kempen (Hrsg.), *Natural Language Generation*, 365–376. Dordrecht: Nijhoff.

Eikmeyer, H.-J. & Schade, U. (1991). Sequenzialization in Connectionist Language Production Models. Erscheint in: *Cognitive Systems.*

Forschergruppe Kohärenz (Hrsg.)(1987). ‚'n Gebilde oder was' – Daten zum Diskurs über Modellwelten. *Kolibri-Arbeitsbericht*, 2. Universität Bielefeld.

Levelt, W.J.M. (1983). Monitoring and Self-Repair in Speech. *Cognition*, 14, 41–104.

Levelt, W.J.M. (1989). *Speaking: From Intention to Articulation.* Cambridge, MA; London: MIT-Press.

MacKay, D.G. (1987). *The Organisation of Perception and Action.* New York, NY: Springer.

Meyer, A.S. (1988). Phonological Encoding in Language Production. Dissertation, Katholieke Universiteit te Nijmegen.

Rumelhart, D.E. & McClelland, J.L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations.* Cambridge, MA: MIT-Press.

Schade, U. (1988). Ein konnektionistisches Modell für die Satzproduktion. In: J. Kindermann & C. Lischka (Hrsg.), *Workshop Konnektionismus*, Arbeitspapiere der GMD, 329, 207–220. St. Augustin.

Schade, U. (1990a). Konnektionistische Modellierung der Sprachproduktion. Dissertation, Universität Bielefeld.

Schade, U. (1990b). Kohärenz und Monitor in konnektionistischen Sprachproduktionsmodellen. In: G. Dorffner (Hrsg.), *Konnektionismus in Artifical Intelligence und Kognitionsforschung. 6. Österreichische Artificial-Intelligence-Tagung (KONNAI): Proceedings*, 18–27. Berlin, Heidelberg: Springer.

Schegloff, E.A. (1979). The Relevance of Repair to Syntax-for-Conversation. In: T. Givón (Hrsg.), *Syntax and Semantics*, Vol. 12. New York, NY: Academic Press.

Schegloff, E.A., Jefferson, G. & H. Sacks (1977). The Preference for Self-Correction in the Organization of Repair in Conversation. *Language*, 2, 361–382.

Stemberger, J.P. (1985a). An Interactive Activation Model of Language Production. In: A.W. Ellis (Hrsg.), *Progress in the Psychology of Language*, Vol. 1, 143–186. London: Erlbaum.

Stemberger, J.P. (1985b). *The Lexicon in a Model of Language Production.* New York, NY: Garland Publishing.

Stemberger, J.P. (1990). Wordshape Errors in Language Production. *Cognition*, 35, 123–157.

Stefan Schierholz / Eric Windisch
IWBS - IBM Heidelberg
Wilckenstr. 1a
6900 Heidelberg

Quantitative Analysen von Bedeutungswörterbüchern

1. Das Untersuchungsziel

Maschinenlesbare Wörterbücher sind zur Zeit ein zentrales Thema in der Computerlinguistik, KI-Forschung und Lexikographie. Insbesondere im englischsprachigen Raum existiert eine Vielzahl maschinenlesbarer Wörterbücher, die nicht nur speziellen Anwendungen dienen, sondern von Jedermann (z.B. auf CD-ROM) käuflich erworben werden können. Bei diesen Lexika werden algorithmische Wörterbuchanalysen nicht nur zum Zwecke der Wörterbuchoptimierung, sondern auch zur Erstellung von lexikalischen Wissenbasen angestellt. Diese lassen sich anschließend in Sprachverarbeitungssystemen mit unterschiedlichen Zielsetzungen einsetzen.
Im Deutschen gibt es meist nur maschinenlesbare Wörterbücher, die entweder sehr klein oder nur für konkrete Anwendungen erstellt sind. Insbesondere unter den Bedeutungswörterbüchern fehlen solche, die maschinenlesbar sind und auch für linguistische Forschungen zur Verfügung stehen.
Im folgenden wird gezeigt, daß auch ohne maschinenlesbare Formate umfangreiche empirische Analysen zu Bedeutungswörterbüchern durchgeführt werden können. Im Mittelpunkt des Untersuchungsinteresses stehen lexikologische und lexikographische Analysen, in denen für die theoretische und praktische Lexikographie verwertbare Resultate erlangt werden.
Mit Hilfe quantitativer Methoden werden systematisch erhobene Datencorpora ausgewertet. Damit soll die Untersuchungsmethode der Fallbeispielsammlung zur Stützung linguistischer Hypothesen ausgeschlossen sein. Weiterhin wird gezeigt werden, daß quantitativ orientierte Arbeiten nicht nur per se aussagekräftig, sondern eine wichtige Stütze qualitativ arbeitender lexikographischer Forschung darstellen.
Untersuchungsgegenstand ist das Duden-Universalwörterbuch in den Ausgaben von 1983 (= DUDUNI) und 1989 (= DDUW).
Bearbeitet werden die folgenden Fragestellungen:

1. Ist der Wortschatz der deutschen Gegenwartssprache im DDUW angemessen lemmatisiert ?
2. Bleibt die Qualität der Bedeutungserklärungen über das gesamte Wörterbuch hinweg konstant ?

Alle Untersuchungsschritte werden durch Tabellen und Abbildungen kleinschrittig nachvollziehbar gemacht.

2. Der Gang der Untersuchung

In acht Bedeutungswörterbüchern des Deutschen sind die Seitenmengen pro Initial gezählt worden. Damit werden die Raumanteile, die zu jedem Anfangsbuchstaben subsumiert sind, erfaßt. Zur unmittelbaren Vergleichbarkeit werden die Raumanteile in Prozent umgerechnet. Für sechs Wörterbücher (DTV-Wahrig 1981, Deutsches Wörterbuch 1988, Handwörterbuch der Deutschen Gegenwartssprache 1984, Großes Wörterbuch der Deutschen Sprache 1976-1981, Brockhaus-Wahrig 1980-1984, Wörterbuch der Deutschen Gegenwartssprache 1965-1977) wird der Mittelwert des Raumanteils pro Initial errechnet, so daß die Abweichungen der Duden-Universalwörterbuchanteile vom Durchschnitt erkennbar werden.
Die Daten der beiden Ausgaben des Universalwörterbuchs weisen keine nennenswerten Abweichungen untereinander auf. Gegenüber dem Durchschnitt der anderen Wörterbücher enthalten DUDUNI und DDUW von "A" bis "H" fast nur geringere Raumanteile, während die Seitenmengen ab "I" mit Ausnahme von "U" und "V" jeweils größer sind. Die geringsten Anteile findet man im DUDUNI/DDUW unter "B", "E", "F" und "G", dagegen aber besonders hohe Werte zu "P" und "T".
Mit diesen Datenanalysen kann aber nicht geklärt werden, ob zu den genannten Initialen die Wörterbuchartikel extrem lang bzw. kurz sind oder ob das enthaltene Lemmainventar in den Universalwörterbüchern besonders vom Durchschnitt abweicht.
In einem zweiten Untersuchungsschritt werden die absoluten Lemmamengen des DDUW pro Initial gezählt und in Beziehung zu den Seitenanteilen gesetzt. Dadurch läßt sich zu jedem Initial die Lemmadichte pro Seite berechnen. Je höher die Lemmadichte pro Initial desto kürzer ist die Wörterbuchartikellänge und desto weniger ausgiebig ist die lexikographische Bearbeitung ausgefallen. Unter den Anfangsbuchstaben "A", "H", "I" und "Z" findet man extrem wenige Lemmata pro Seite, während vor allem zu "F", "M", "N", "O", "P" und "U" die Lemmadichte erheblich höher als der errechnete Mittelwert von 70.57 Lemmata pro Seite liegt. Im Zusammenhang mit den Raumanteilen pro Initial, die mit den durchschnittlichen Raumanteilen in deutschen Bedeutungswörterbüchern verglichen worden sind, lassen sich für einige Initialen jetzt dezidierte Aussagen über den Grund ihrer Abweichungen von den Mittelwerten machen.
Der dritte Untersuchungsschritt bezieht Textcorpora in die vergleichenden Analysen ein. Dies erscheint notwendig, weil bislang nur wörterbuchinterne Vergleiche angestellt worden sind und somit Fehlerquellen, die sich in Wörterbüchern systematisch wiederholen, nicht berücksichtigt werden können. Zu verschiedenen Textcorpora aus Zeitungstexten wird der durchschnittliche prozentuale Anteil jedes Anfangsbuchstaben an allen verschiedenen Wortformen errechnet. Der Vergleich zwischen den Lemmata (DDUW) und Wortformen (Texte) ist unproblematisch, wenn man die Vergleiche auf die Initialenverteilung beschränkt. Anhand der Abweichungen der DDUW-Lemmaverteilung pro Initial von den Textwerten kann man erkennen, in welchen Bereichen besonders viele bzw. wenige Lemmalücken im DDUW existieren.
Durch eine Zusammenstellung aller erhobenen Resultate kann man jetzt für jedes Initial im DDUW ziemlich genau angeben, ob die Abweichungen zu Vergleichswerten aufgrund von Lemmalücken oder weniger umfangreicher lexikographischer Bearbeitung bestehen. Ebenso gibt es natürlich Werte, die auf eine überdurchschnittlich gute Bearbeitung oder ein überproportinal umfangreiches Lemmainventar hinweisen. So werden zum Buchstaben "A" verhältnismäßig viele Lemmalücken bei umfangreicher lexikographischer Bearbeitung errechnet. Zu "U" gibt es viele Lemmalücken und extrem kurze Wörterbuchartikel. "H" und "P" enthalten relativ viele Lemmata, aber bei "H" sind die Wörterbuchartikel länger als im DDUW-Durchschnitt, während zu "P" die Lemmadichte pro Seite sehr hoch ist.
Die Resultate zeigen Unregelmäßigkeiten in der Initialverteilung zum DDUW. Sprachliche Merkmale des Deutschen (Präfigierung, Komposita, Fremdwortanteile, Konversion), die diese Unregelmäßigkeiten verursachen, sind in die Überlegungen einbezogen, sollten aber Gegenstand weiterer Diskussionen und empirischer Analysen sein.
Die Ergebnisse zeigen jedoch auch, daß quantitative Analysen eine wesentliche Voraussetzung und wichtige Hilfestellung für qualitative lexikographische Analysen darstellen. Wörterbuchoptimierungen können zukünftig gezielter vorgenommen werden, lexikologische Analysen anhand des DDUW sollten unter Berücksichtigung der dargestellten Resultate angestellt werden.
Außerdem kann gezeigt werden, daß makrostrukturelle Wörterbuchanalysen auch ohne maschinenlesbare Formate durchführbar sind. Trotzdem ist für die deutschsprachige Lexikographie das Vorhandensein maschinenlesbarer Wörterbücher ein Desiderat, dem nicht genug Nachdruck verliehen werden kann. Angesichts der fortgeschrittenen Entwicklung in anderen Ländern verbleibt die deutsche Wörterbuchlandschaft nach wie vor im Zustand eines lexikographischen Entwicklungslandes.

Literaturliste (Bedeutungswörterbücher):

Duden. Das große Wörterbuch der deutschen Sprache in sechs Bänden. Herausgegeben und bearbeitet vom Wissenschaftlichen Rat und den Mitarbeitern der Dudenradaktion unter der Leitung von Günther Dosdrowski. Mannheim/Wien/Zürich 1976-1981.

Duden. Deutsches Universalwörterbuch. Hrsg. und bearb. vom Wissenschaftlichen Rat und den Mitarbeitern der Dudenredaktion unter Leitung von Günther Drosdowski. Mannheim/Wien/Zürich 1983.

Duden. Deutsches Universalwörterbuch. Hrsg. und bearb. vom Wissenschaftlichen Rat und den Mitarbeitern der Dudenredaktion unter der Leitung von Günther Drosdowski. 2., völlig neu bearbeitete und stark erweiterte Auflage. Mannheim/Wien/Zürich 1989.

Kempcke, Günther (Autorenkollektiv): Handwörterbuch der deutschen Gegenwartssprache in zwei Bänden. Berlin 1984.

Klappenbach, Ruth/Steinitz, Wolfgang (Hrsg.): Wörterbuch der deutschen Gegenwartssprache. 2., durchgesehene Aufl. Berlin 1965-1977.

Wahrig, Gerhard (Hrsg.): Brockhaus-Wahrig. Deutsches Wörterbuch in sechs Bänden. Wiesbaden/Stuttgart 1980-1984.

Wahrig, Gerhard (Hrsg.): dtv-Wörterbuch der deutschen Sprache. Hrsg. in Zusammenarbeit mit zahlreichen Wissenschaftlern und anderen Fachleuten. 4. Aufl. München 1981.

Wahrig, Gerhard (Hrsg.): Deutsches Wörterbuch. Herausgegeben in Zusammenarbeit mit zahlreichen Wissenschaftlern und anderen Fachleuten. Völlig überarbeitete Neuausgabe. München 1988.

The Measurement of Morphosyntactic Properties: A First Attempt

Peter Schmidt
Universität Konstanz, Fachgruppe Sprachwissenschaft - Slavistik -,
Postfach 5560, 7750 KONSTANZ

The present contribution constitutes a first attempt at a quantitative analysis of morphosyntactic phenomena in natural languages, which is based on the transfer and analogical application of concepts and methods that were developed for and successfully applied to the measurement of phonological and morphological properties; cf. ALTMANN/LEHFELDT(1973), ALTMANN/LEHFELDT(1980), LEHFELDT(1985).

The results of this study are part of extended research on general questions of morphosyntax and on the morphosyntax of Modern Russian conducted under the auspices of the DFG at the University of Constance, Chair of Slavic Linguistics, and directed by Prof. Dr. W.Lehfeldt; cf. SCHMIDT/LEHFELDT(in preparation).

The present analysis presupposes for a given natural language a specification of (i) **w**(ord)**c**(lasses), defined as sets of **w**(ord)**f**(orms), (ii) the **g**(rammatical)**c**(ategories) of each word class, (iii) the **g**(ra)**m**(memes) of each grammatical category, (iv) **l**(exeme)**c**(lasses), i.e., the sets of all **lex**(emes) of a given word class, and (v) an inventory of **s**(urface)**s**(yntactic labelled dependency)**r**(elations) of the type assumed in the "Meaning ⇔Text" approach of MEL'ČUK(1979).

A grammatically analyzed (binary) syntagm may now be defined as an ordered triple $\langle$(governing)$\mathbf{wf_i}$,$\mathbf{ssr_r}$,(dependent)$\mathbf{wf_k}\rangle$, $\mathbf{wf_i} \in \mathbf{wc_l}$, $\mathbf{wf_k} \in \mathbf{wc_m}$: $\mathbf{wf_i} \xrightarrow{r} \mathbf{wf_k}$, where each $\mathbf{wf_n}$ consists of a $\mathbf{lex_o}$ and a **gr**(ammemic component - which is empty in the case of uninflected forms) of $\mathbf{p\ gm_q}$, containing one grammeme per grammatical category of the respective word class: $\mathbf{wf_n} = \langle \mathbf{lex_o}, \mathbf{gr_s} \rangle = \langle \mathbf{lex_o}, \langle \mathbf{gm_1}, ..., \mathbf{gm_p} \rangle \rangle$.

The German attributive syntagm *alter Hut*('old hat', nominative singular masculine, syntactic construction: $\mathbf{N} \xrightarrow{attr} \mathbf{A}$) may serve as an illustration: $\langle$HUT,$\langle$**nom,sg,m**$\rangle\rangle \xrightarrow{attr} \langle$ALT,$\langle$**nom,sg,m,pos(itive)**$\rangle\rangle$

The *signifiant* of a syntagm may be defined, disregarding word order, prosodic means etc. for the moment, as a binary set (or, more exactly, a binary multiset: $\{x,x\} \neq \{x\}$) $\{\mathbf{seg_i}, \mathbf{seg_k}\}$ of **seg**(ments), with $\mathbf{seg_i}$ expressing $\mathbf{wf_i}$ and $\mathbf{seg_k}$ expressing $\mathbf{wf_k}$: {*alter*,*Hut*}.

A (binary minimal) s(yntactic)c(onstruction) may thus be regarded as the set containing all well-formed analyzed syntagms (or, on the expression plane, as the set of their binary sets of segments) with identical governing word class, identical dependent word class, and identical surface-syntactic dependency relation. It is technically convenient here to suppress mention of the surface syntactic relation of a syntagm and refer to the corresponding ordered pair of analyzed word forms instead: $\langle\langle$HUT,$\langle$**nom,sg,m**$\rangle\rangle$, $\langle$ALT,$\langle$**nom,sg,m,pos**$\rangle\rangle\rangle$.

This allows us to represent a syntactic construction by that (potentially, but not typically, improper) subset of the cartesian product of its governing and its dependent word class which contains all word form pairs figuring in well-formed syntagms of the respective syntactic construction. Thus, for the German attributive construction $\mathbf{N} \xrightarrow{attr} \mathbf{A}$ it is the case that, for example,

$\langle\langle$WEIN,$\langle$**gen,sg,m**$\rangle\rangle$,$\langle$ALT,$\langle$**gen,sg,m,pos**$\rangle\rangle\rangle \in \mathbf{N} \xrightarrow{attr} \mathbf{A}$ (*alten Weines* 'old wine', genitive),

$\langle\langle$FLASCHE,$\langle$**nom,pl,f**$\rangle\rangle$,$\langle$NEU,$\langle$**nom,pl,f,sup**(erlative)$\rangle\rangle\rangle \in$
$\mathbf{N} \xrightarrow{attr} \mathbf{A}$ (*neueste Flaschen* 'newest (i.e., latest) bottles'), but

$\langle\langle$HUT,$\langle$**nom,pl,m**$\rangle\rangle$,$\langle$ALT,$\langle$**dat,sg,m,sup**$\rangle\rangle\rangle \notin \mathbf{N} \xrightarrow{attr} \mathbf{A}$
(**ältestem Hüte* 'oldest (dative singular) hats (nominative plural)'),

The cartesian product of governing and of dependent word class, which delimits the 'event space' for the formation of syntactic constructions with a given governing word class and a given dependent word class as its 'ingredients', can be spelled out in more detail as

$$(\mathbf{lc\ x\ gc_1\ x\ ...\ x\ gc_p})_{\text{governor}}\ \mathbf{x}\ (\mathbf{lc\ x\ gc_1\ x\ ...\ x\ gc_q})_{\text{dependent}}.$$

As there exists a corresponding subset of this cartesian product for every syntactic construction, the various morphosyntactic properties of individual syntactic constructions as well as similarities of different syntactic constructions with respect to given (combinations of) morphosyntactic properties can be modelled by corresponding set-theoretic properties and relations defined on this space.

Types of morphosyntactic marking of syntactic relations may thus be defined in a straightforward manner, on the basis of the structural characteristics of the associated sets of syntagms, or sets of word form pairs, respectively; they can be explicated as combination restrictions on certain components of the corresponding cartesian product. To give just two purely illustrative examples: grammatical agreement is reflected in the exclusion of non-identical combinations of certain identical grammatical categories of governor and dependent, whereas (morphosyntactic) government may be defined as selectional interdependence of syntactically governing lexemes and case grammemes of the syntactic dependent, etc.

Given these prerequisites, one can define and measure diverse properties of syntactic constructions (both on the level of analyzed syntagms

and on the level of their *signifiants*) which reflect different aspects of the relevance of morphosyntactic marking in syntactic analysis/parsing, e.g. (definitions are primarily illustrative and cover only the simplest conceivable cases):

(i) **selectivity** of a syntactic construction: A syntactic construction shall be called (holistically) **selective** to the extent that the set of its well-formed syntagms is a proper subset of the cartesian product of its word classes. Selectivity (or its complement, 'permissivity') of a syntactic construction may be taken as an indicator of the (abstract) ease/difficulty (compared to other syntactic constructions with the same governing and dependent word classes) to find word form pairs in a string that are potential instances of the syntactic construction in question.

Given now a governing word class $\mathbf{wc_j}$ and a dependent word class $\mathbf{wc_k}$ of a $\mathbf{sc_i}$, with cardinalities $|\mathbf{wc_j}|$, $|\mathbf{wc_k}|$, one possible measure of holistic selectivity is

$$\text{sel}(\mathbf{sc_i}) = 1 - \frac{\text{number of well-formed syntagms of } \mathbf{sc_i}}{|\mathbf{wc_j}| \times |\mathbf{wc_k}|} \quad (0;\frac{1}{|\mathbf{wc_j}| \times |\mathbf{wc_k}|});$$

$$\Rightarrow \text{sel}'(\mathbf{sc_i}) = \frac{\text{sel}(\mathbf{sc_i})}{\max(\text{sel}(\mathbf{sc_i}))} \quad (0;1),$$

with $\max(\text{sel}(\mathbf{sc_i}))$ being the maximum value of the original measure $\text{sel}(\mathbf{sc_i})$.

This measure may be particularized for individual factors or combinations of factors, such as **morphosyntactic/agreement/government/lexical/... selectivity**.

(ii) **relative weight of particular types of morphosyntactic marking** in the overall (morphosyntactic) marking of a syntactic construction.

(iii) **g**(overnor)-**d**(ependent)-(or, **d-g-**)**predictivity**: A syntactic construction shall be called **g-d-(d-g-)predictive** to the extent that a given governor (dependent) - or certain given (combinations of) features of the governor (dependent) - narrow(s) down the choice of dependents (governors) - or of certain (combinations of) features of the dependents (governors). The degree of predictivity of a syntactic construction models a property relevant to syntactic analysis, in that it reflects the extent to which knowledge of one of the (potential) members of a binary syntagm restricts the features of potential syntactic mates.

The **overall g-d-predictivity** of a syntactic construction $\mathbf{sc_i}$ can be defined as the average degree of predictability of the dependent word forms of $\mathbf{sc_i}$ on the basis of the governing word forms of $\mathbf{sc_i}$, with the latter being definable for individual governing word forms via the number of permitted dependent $\mathbf{wf_k}$ for a given governing $\mathbf{wf_j}$: Let $\mathbf{p}$ $(1;\mathbf{k}\ (\leq |\mathbf{wc_d}|))$ be the number of dependent word forms of $\mathbf{sc_i}$ that are well-formed for a given governing word form $\mathbf{wf_j}$, and let $\mathbf{k}$ be the total amount of well-formed dependent word forms of $\mathbf{sc_i}$.

Then $\text{g-d-pred}(\mathbf{wf_j},\mathbf{wc_d}) = \frac{\mathbf{k} - \mathbf{p}}{\max(1,\mathbf{k} - 1)}$ (0;1), and

$$\text{g-d-pred}(\mathbf{sc_i}) = \frac{\sum_{i=1}^{n} \text{g-d-pred}(\mathbf{wf_i},\mathbf{wc_d})}{\mathbf{n}} \quad (0;1),$$

$\mathbf{n}$ the number of well-formed governing word forms of $\mathbf{sc_i}$.

This measure reflects only those restrictions of the dependent (governor) which are conditioned by variation of features of the governor (dependent) - within the limits of its well-formedness - and deliberately excludes those restrictions of the dependent (governor) that are independent of the choice of the governor (dependent). If these 'pseudo-predictions' were to be included in order to reflect all types of restrictions of the dependent irrespective of their source, $\mathbf{k}$ in the above formula would have to be replaced by $|\mathbf{wc_d}|$.

Predictivity can again be particularized for individual factors or combinations of factors. Thus, **(g-d-) government predictivity** (in the sense of morphosyntactic government assumed above) may be defined as $\text{g-d-pred}(\mathbf{wf_j},\mathbf{case_d})$.

(iv) Various types of **homonymy/ambiguity**, which can again be measured for whole syntactic constructions or for specific (combinations of) factors, e.g.

- **intra-construction homonymy**, i.e., homonymy of syntagms belonging to one and the same syntactic construction: This can be measured in the simplest case (no synonymous expressions of syntagms) by the ratio of the number of binary sets of segments to the number of analyzed word form pairs of a syntactic construction. Let $\text{sp}(\mathbf{sc_i})$ be the set of binary sets of segments, $\text{wp}(\mathbf{sc_i})$ the set of analyzed word form pairs of $\mathbf{sc_i}$, then

$$\text{hom}_{\text{intra}}(\mathbf{sc_i}) = 1 - \frac{|\text{sp}(\mathbf{sc_i})|}{|\text{wp}(\mathbf{sc_i})|} \quad (0;1 - 1/|\text{wp}(\mathbf{sc_i})|), \Rightarrow$$

$$\text{hom}_{\text{intra}}'(\mathbf{sc_i}) = \frac{\text{hom}_{\text{intra}}(\mathbf{sc_i})}{\max(\text{hom}_{\text{intra}}(\mathbf{sc_i}))} \quad (0;1).$$

- **inter-construction homonymy** of two syntactic constructions: This may be measured in the simplest case (no ambiguities within each of the syntactic constructions) by the ratio of the cardinality of the intersection of the sets of binary sets of segments of the two constructions to its possible maximum value, viz. the cardinality of the smaller of the two sets:

$$\text{hom}_{\text{inter}}(\mathbf{sc_i},\mathbf{sc_j}) = \frac{|\text{sp}(\mathbf{sc_i}) \cap \text{sp}(\mathbf{sc_j})|}{\min(|\text{sp}(\mathbf{sc_i})|,|\text{sp}(\mathbf{sc_j})|)} \quad (0;1).$$

(v) g(overnor)-/**d**(ependent)-**marking** of syntactic relations; cf. NICHOLS(1986).

On the basis of the measures for individual syntactic constructions, these and other morphosyntactic properties may also be measured for arbitrary subclasses of syntactic constructions or for complete languages, thus providing for a holistic characterization and comparison of languages with respect to their morphosyntactic properties. **Selectivity of a language $\mathbf{l_m}$**, e.g., may be defined as

$$\text{sel}(\mathbf{l_m}) = \frac{\sum_{i=1}^{n} \text{sel}(\mathbf{sc_i})}{\mathbf{n}} \quad (0;1), \ \mathbf{n} \text{ the number of syntactic constructions in } \mathbf{l_m}.$$

The illustrated measures of various morphosyntactic properties that are relevant to syntactic analysis could be applied to the study of the relative importance and the interrelations of the various existing types of morphosyntactic marking in natural languages. When supplemented by analogous (easily devisable) measures of the disambiguating function of word order and lexical selection restrictions, they could be employed in the comparative study of the relative weight of these factors in (human or mechanical) syntactic disambiguation (and syntactic decoding in general), and of their typological interrelations. Furthermore, most of the measures cited may easily be adapted to the measurement of analogous properties in other linguistic subsystems, such as inflectional morphology or the lexicon, since they reflect general semiotic, rather than exclusively morphosyntactic, properties of natural language.

References:

- ALTMANN, G./LEHFELDT, W.: *Allgemeine Sprachtypologie. Prinzipien und Meßverfahren.* München 1973
- ALTMANN, G./LEHFELDT, W.: *Quantitative Phonologie.* Bochum 1980
- LEHFELDT, W.: *Спряжение украинского глагола* . München 1985
- MEL'ČUK, I.A.: *Dependency Syntax.* Ann Arbor 1979
- NICHOLS, J.: *Head-marking and dependent-marking grammar.* Language **62** (1986), 56-119
- SCHMIDT, P./LEHFELDT, W.: *Die zweigliedrigen Wortfügungen (словосочетания) des Russischen.* München: O.Sagner (in preparation)

Walter F. Sendlmeier

PHONETISCH/PHONOLOGISCHE AUSBALANCIERTHEIT VON TESTMATERIALIEN ZUR SPRACHGÜTEBEURTEILUNG

Zusammenfassung: Die Einzellauthäufigkeitsverteilung gilt noch immer als Kardinalkriterium für die phonetisch/phonologische Ausbalanciertheit von sprachlichen Testmaterialien in der Sprachgütemessung und in der Sprachaudiometrie. Es wird aufgezeigt, inwiefern dieser Zugang aus phonetischer Sicht als sehr einseitig bezeichnet werden muß. Eine Reihe von Parametern wird vorgeschlagen, die in stärkerem Maße geeignet erscheinen, die tatsächliche phonetische Struktur einer Sprache bei der Gestaltung von Testlisten abzubilden.

Die bisherigen Testmaterialien für die Sprachgütebeurteilung und die Sprachaudiometrie lassen sich nach einigen allegemeinen Merkmalen charakterisieren. Bei den Vertretern verschiedener Testformen zur isolierten Wort- oder Lauterkennung, die im europäischen oder nordamerikanischen Raum verwendet werden, findet man hinsichtlich der Antwortform zwei Grundtypen, nämlich einmal die offene Antwortform, wie sie etwa im Freiburger Test oder in den Harvard PB-50-Listen und deren Weiterentwicklung als CID W-22-Listen vorliegen, und die geschlossene Antwortform, als deren Vertreter etwa der "Modified Rhyme Test" (MRT) oder für das Deutsche der Sotschek-Reimtest zu nennen sind.

In den meisten Tests beider Antworttypen wird mit Einsilbern gearbeitet, wobei die Silbenstruktur in der Regel sogar auf die Form Konsonant-Vokal-Konsonant (oft als K-V-K oder C-V-C abgekürzt) eingeschränkt wird. Die möglichst gute Übereinstimmung der Einzellauthäufigkeit von Testmaterialien und großen Korpora der Bezugssprachen gilt heute als zentrales Kriterium für die phonetisch/-phonologische Repräsentativität der jeweiligen Testlisten, die für die Überprüfung des Hörverstehens von Lautsprache erstellt wurden. Darüberhinaus wird bei der Entwicklung eine interne phonetisch/-phonologische Ausbalanciertheit zwischen den einzelnen Sublisten angestrebt.

Neben dem Kardinalkriterium der Ausgewogenheit der Einzellauthäufigkeiten wurden andere Aspekte bei der Entwicklung von Testmaterialien völlig übersehen oder für weniger wichtig erachtet. An diesem Vorgehen muß aus phonetisch/phonologischer Sicht deutliche Kritik geübt werden. Selbst wenn man Einzellauthäufigkeiten als entscheidendes Kriterium akzeptierte, bliebe das Bemühen, diese ausschließlich auf CVC-Einheiten zu projizieren, in hohem Maße fragwürdig. Dies wird deutlich, wenn man das Vorgehen genauer betrachtet, nach dem versucht wird, die Häufigkeitsverteilungen der Einzellaute im Testmaterial an die Verteilung der Laute in der jeweils zu testenden Sprache anzupassen.

Menzerath kommt bei seiner Klassifikation nach den Kriterien der Silbenzahl, der Lautzahl und der Lautfolge zu Ergebnissen, die zeigen, daß die Einsilber von der Struktur CVC im Typenhäufigkeitsgebirge des deutschen Wortschatzes eine Randstellung einnehmen. Für den vorliegenden Beitrag wurde ausgehend von den zehntausend häufigsten Wörtern des recht aktuellen LIMAS-Korpus die Verteilung der Worthäufigkeiten entsprechend ihrer Silbenstruktur und Silbenzahl erneut analysiert.

In der Entwicklung von Testmaterialien ist es ratsam, einen Kompromiß zwischen Type- und Token-Zählung anzustreben: Rein formale Zählungen auf der Basis von Token-Auswertungen führen aufgrund des extrem häufigen Auftretens von Funktionswörtern zu Verzerrungen in Einsilbertests. Die völlige Vernachlässigung von Gebrauchshäufigkeiten ist jedoch keine akzeptable Alternative. Ebenfalls scheint es ratsam, mehrere Parameter zu berücksichtigen, die gleichermaßen die phonetische Struktur einer Sprache bestimmen. Hier sind neben den Einzellauthäufigkeiten die Häufigkeitsverteilungen hinsichtlich der Silbenzahl, der Silbenstruktur und der Wörter sowie der Lautübergänge von Bedeutung. Bei der Erstellung von Testmaterialien stellt sich darüberhinaus stets die Aufgabe, eine Balance zwischen der Forderung nach ökologischer Validität einerseits und Reliabilität andererseits zu finden. Durch die Reduktion der lautlichen Einheiten und ein Reim-Test-Design werden Worthäufigkeitseffekte und individuelle Sprachgewohnheiten als Ursache für eine Variation der Testergebnisse weitgehend ausgeschlossen.

Sendlmeier, W.F. (1990): "Testmaterial Sprache - grundsätzliche Überlegungen zu Verfahren der Sprachgütemessung und Sprachaudiometrie", in: P. Plath (Ed.): Neue Technologien in der Hörgeräteakustik - Herausforderungen an die Audiologie. 5. Multidisziplinäres Colloquium der Geers-Stiftung. Bonn, 42-65.

Sendlmeier, W.F. (1991): "Wie testet man Hörverstehen? Eine kritische Analyse sprachaudiometrischer Testverfahren", in: W. Sendlmeier & W. Hess (Ed.): Experimentelle und Angewandte Phonetik. Stuttgart (Steiner).

Mildred L. G. Shaw & Brian R. Gaines

A METHODOLOGY FOR ANALYZING TERMINOLOGICAL AND CONCEPTUAL DIFFERENCES IN LANGUAGE USE ACROSS COMMUNITIES

The study of discourse within specialized communities with a common disciplinary, or goal-directed, focus requires methodologies for the empirical derivation and comparative analysis of both the conceptual and terminological structures used by participants. This paper reports on research studies over the past sixteen years in which we have developed a theoretical framework, quantitative empirical methodologies, and computer-based analysis tools, for modeling terminological and conceptual relations in language use across communities. In particular, recent developments are illustrated of on-line, computer-based tools for the interactive elicitation of concepts and terminology using graphic interfaces on personal computers. Some new extensions allowing the use of these tools over networks to support discourse and knowledge processes in dispersed communities are described. The presentation emphazises the close links between the theroretical foundations, linguistic and knowledge representation concepts, empirical methodology, its implementation in computer-based tools, and practical applications. It illustrates the powerful tools now available for the study of quantitative linguistic phenomena in specialized communities.

George Silnitsky (Smolensk)

CORRELATIONAL ANALYSIS AND CLUSTERIZATION OF VERBAL FEATURES IN ENGLISH AND GERMAN

The research program discussed in this paper, realized by a group of linguists of the Smolensk Teachers Training Institute, is empirically based on a full list of English verbs recorded in Hornby's "Advanced Learner's Dictionary" (5684 lexical units). Every verb in the list is characterized on the dichotomic principle by the presence or absence of 119 semantic, syntactic, diachronic, morphologic, derivational, phonetic features.

A correlational analysis of this data base (on Pearson's criterion) elicited the statistically relevant (positive or negative) connections between all the logically possible pairs of verbal features (of the above-mentioned set). The positively correlated features were then grouped into three clusters based on the semantic classes of "energic", "informational" and "ontological" verbal meanings.

The class of ENERGIC verbal meanings reflects various processes of physical energy transformation: locomotion ("move", "throw"), change of form ("bend", "break"), physical processes ("burn", "melt"), physiological processes ("live", "die", "grow").

The class of INFORMATIONAL verbal meanings reflects vaious types of psychical and informational processes and relations: perceptual ("see", "hear"), intellectual ("know", "prove"), emotional ("admire", "astonish"), volitional ("want", "inspire"), semiotic ("mean", "denote", "express").

The ONTOLOGICAL class, which includes verbal meanings of existence ("exist", "create", "destroy"), possession ("have", "give", "lend"), evaluation ("improve", "spoil"), quantity ("increase", "decrease"), temporal characteristics ("prolong", "postpone", "date"), social processes ("industrialize", "rule"), is of a more generalized status than the first two classes. Depending on the concrete lexical environment of the verb, ontological meanings may be realized either in an energic ("to construct a bridge", "to acquire property", "to improve one's health", "to increase one's income") or in an informational ("to construct a theory", "to acquire knowledge", "to improve one's manners") sense.

Energic meanings are positively correlated with the following verbal features (only the crucially diagnostic features are listed below; correlation coefficients given in brackets): OE origin (.06), monosyllabic phonetic structure (.21), monomorphemic structure ("root verbs": .22), intransitivity (.18). They are negatively correlated with affixal derivatedness (-.07) and "neutral" as regards semantic causativity/noncausativity.

Ontological meanings are directly opposed to energic ones in their correlational affiliations, being positively linked with such features as NE origin (.04), polysyllabic structure (.11), polymorphemic structure (.12), derivatedness (.13), causativity (.15), transitivity (.08).

Informational verbal meanings are diachronically indifferent to all the periods of the evolution of English, i.e. they originated more or less uniformly during the whole history of the language. They are positively correlated with various types of syntactic (indirect and prepositional object: .07; object clause: .21; the accusativus-cum-infinitivo construction: .11) and derivational (nominal: .07; adjectival: .07) extraverbal valencies.

The following interpretation of the above-given classification may be proposed.

Energic verbal meanings reflect the main material exigencies of human existence and therefore constitute an indispensable semantic

component of language from the earliest period of its recorded evolution. This early origin of energic verbs and their consecutive longer life-span (only such verbs were considered which have persisted till the twentieth century) determine their tendency toward simplicity of formal and semantic structure; the latter, in its turn, accounts for the positive connection of energic verbs with intransitivity.

In opposition to these, ontological verbal meanings, being of a more abstract nature, tend to a later, derivated origin and consequently to a more complex formal, semantic and syntactic structure.

Informational verbal meanings, in distinction to the first two types, are represented on two semantic levels: that of the psychic process itself and that of its c o n t e n t. On the first criterion the psychical processes of sight, hearing, cognition, volition, etc. pertain to the sphere of basic human properties and therefore tend to be represented in the most primeval strata of language. But from the point of view of their content, with its tendency to an increased complexity in the course of history, informational verbs are prone to appear in the more advanced phases of language evolution. These two opposite tendencies counterbalance each other with the result that informational verbs are diachronically "neutralized". The same semantic factors may be held to account for the heightened syntactic and derivational valency of informational verbs mentioned above.

A preliminary correlational analysis of the same set of verbal features in German (effected on a list of 400 verbs) has brought to light a high degree of proximity between the two languages from the point of view of the criteria under discussion.

Marek Świdziński

Verb patterns in the Polish vocabulary and texts

1. Aim

In this paper I will present a database of 1500 Polish verbs compiled by a group of my students at Warsaw University in 1988-90. Each entry of this database gives, among others, full syntactic information on the lexical unit it describes. The empirical investigations have provided, as a kind of by-product, some statistical data, which, I hope, may be of interest to those who deal with quantitative linguistics.

2. The Universal Basic Dictionary of Contemporary Polish

The work I will refer to in what follows has been undertaken within a project initiated in the early 1980s by a group of linguists and computer specialists at Warsaw University. The project is aimed at the construction of the Universal Basic Dictionary of Contemporary Polish (henceforth: UBDCP), intended as a small-size grammar- oriented guide to the Polish lexicon (see Saloni et al. (1990)). Up to now, some types of grammatical information for the UBDCP have been completely worked out1. They are also important to a lexicographer who has usually no intuition concerning the systemic or textual frequency of various grammatical phenomena.

3. The notion of sentence schema

In any dictionary, particular attention should be paid to verbs, as the verb constitutes the core of a sentence, of its meaning and its grammatical structure. Therefore, each verbal entry in the UBDCP will include a set of sentence schemata which it founds. Unfortunately, Polish dictionaries traditionally neglect the syntactic features of the lexical units, which makes them practically useless to non-native speakers2.

The UBDCP is based on two formalized descriptions of Polish, given in Saloni and Świdziński (1987), Szpakowicz (1986), and Świdziński (1991). It is claimed there that there are two classes of sentences in any language (see Saloni and Świdziński (1987: Ch. II). Natural language sentences either

(i) consist of the finite phrase plus (optionally) other phrases required by it (simple sentences3), or
(ii) contain two sentences joined by a coordinate conjunction (compound sentences).

A sentence schema is understood here as a representation of a class of empirical sentences of type (i) by a set of phrases, including the finite phrase as its structural center and its requirees. The finite phrase is a distributional equivalent of the personal wordform of a verb4; the required phrase is what Tesniere named actant (Tesnie're (1957)), i. e., a syntactic position, rather than a morphologically defined textual unit.

4. The list of syntactic schemata for Polish

In Saloni and Świdziński (1987: Ch XII), a set of sentence schemata for Polish verbs is given. They are divided into two subsets. To one subset, sentence schemata belong that are based on the finite wordform of a standard verb (subject-containing schemata). The other subset contains schemata founded by the finite wordform of a quasi-verb, i. e., of an impersonal verb (subject- less schemata). Both subsets are classified further according to the number and type of the phrases required. Besides the finite phrase, 7 types of phrases are distinguished: nominal, preposi- tional-nominal, adjectival, prepositional-adjectival, infinitival, adverbial, and sentence-like phrases. The set of required phrases contains 0 through 3 units, which gives a repertory of not less than 31 schemata.

The sentence schemata for Polish are the following:

V Verbal schemata (subject-containing schemata)

Nominal phrase in Nominative plus:

V-0: 0 (=nothing)
Jan śpi. 'John is sleeping.'

V-11 Nominal phrase
Jan kupuje dom. 'J. is buying a house.'
V-12 Prepositional-nominal phrase
Jan śmieje się z Marii. 'J. is laughing at Marie.'
V-13 Adjectival phrase
Jan jest głupi. 'J. is stupid.'
V-14 Prepositional-adjectival phrase
Jan wygląda na zm~czonego. 'J. looks tired.'
V-15 Adverbial phrase
Jan zachowuje się niewłaściwie. 'J. behaves improperly.'
V-16 Infinitival phrase
Jan chce spać. 'J. wants to sleep.'
V-17 Sentence-like phrase (=clause)
J. pytą kt"ra godzina. 'J. is asking what time it is.'
V-21 Nominal phrase + Nominal phrase
Jan pożycza Marii książki. 'J. lends M. books.'
V-22 Nominal phrase + Prepositional-nominal phrase
Jan pożycza książki od Marii. 'J. borrows books from M.'
V-23 Nominal phrase + Adjectival phrase
Jan wydaje się Marii miły. 'J. seems nice to M.'
V-24 Nominal phrase + Adjectival phrase (with gender/number agreement)
Jan pamięta Marię młodą. 'J. remembers M. young.'
V-25 Nominal phrase + Prepositional-adjectival phrase
Jan wygląda nam na zmęczonego. 'J. seems tired to us.'
V-26 Nominal phrase + Prepositional-adjectival phrase (with gender/number agreement)
Jan bierze Marię za wykształconą. 'J. takes M. for educated.
V-27 Nominal phrase + Adverbial phrase
Jan stawia kubek tutaj. 'J. puts the cup here.'

V-28 Nominal phrase + Infinitival phrase
Jan każe Marii czekać. 'J. tells M. to wait.'
V-29 Nominal phrase + Sentence-like phrase (=clause)
Jan mówi Marii, że nie ma czasu. 'J. tells M. that he has no time.'
V-210 Prepositional-nominal phrase + Prepositional-nominal phrase
Jan dowiaduje się o Piotrze od Marii. 'J. learns about Peter from M.'
V-211 Prepositional-nominal phrase + Adverbial phrase
Jan mówi o Marii nieżyczliwie. 'J. speaks about M. unfavourably.'
V-212 Prepositional-nominal phrase + Sentence-like phrase (=clause)
Jan wie od Marii, gdzie Piotr mieszka. 'J. has it from M. where P. lives.'
V-213 Adverbial phrase + Adverbial phrase
Jan leci stąd do Berlina. 'J. is flying from here to Berlin.'

Q Quasi-verbal schemata (subjectless schemata)

Q-0 0 (=nothing)
Świta. 'It's dawning.'

Q-11 Nominal phrase
Jana mdli. 'J. feels sick.'
Q-12 Prepositional-nominal phrase
Czas na obiad. 'It's time for dinner.'
Q-13 Adverbial phrase
Jest zimno. 'It's cold.'
Q-14 Infinitival phrase
Trzeba pracować. 'It's necessary to work.'
Q-15 Sentence-like phrase (=clause)
Wiadomo, co Jan powie. 'It's known what J. will say.'
Q-21 Nominal phrase + Nominal phrase
Janowi brak pieniędzy. 'J. is short of cash.'
Q-22 Nominal phrase + Prepositional-nominal phrase
Jana ciągnie do picia. 'J. is tempted to drink.'
Q-23 Nominal phrase + Adverbial phrase
Janowi idzie dobrze. 'J. is doing well.'
Q-24 Nominal phrase + Infinitival phrase
Pracować jest niepodobieństwem. 'It's impossible to work.'
Q-25 Nominal phrase + Sentence-like phrase (=clause)
Janowi żal, że Maria uciekła. 'J. regrets that M. escaped.'
Q-26 Infinitival phrase + Adverbial phrase
Wygrywać jest łatwo. 'It's easy to win.'

5. The database of Polish verbs and quasi-verbs

The database of Polish verbs and quasi-verbs contains 1494 lexical entries. They are all taken from a pocket-size Polish- English dictionary, compiled by K. Billip and Z. Chociłowska (Słownik-minimum (1988)), which makes this selection natural and representative6. All these items have been carefully examined, and sets of sentence schemata have been assigned to each of them, on the basis of the illustrative material given in the respective verbal entries of a general-purpose Polish dictionary (SJP PWN (1978-82)). Słownik generatywny (1980-) was another source of syntactic information, although of minor importance.

Yet another aspect of these calculations seems to be of importance from the quantitativists' point of view: that connected with the only numerical field in the database. The entry of this database is attributed a figure that states how many sentence schemata the verb at question is assigned. The figures are the following:

1: 225 (15.06%); 2: 429 (28.71%); 3: 339 (22.69%); 4: 228 (15.26%); 5: 136 (9.10%); 6: 79 (5.29%); 7: 32 (2.14%); 8: 12 (0.80%); 9: 6 (0.40%); 10: 4 (0.27%); 11: 2 (0.13%).

One can see that most verbs fulfil 1 to 3 sentence schemata (993, i. e., 66.47%); of these verbs, 429 (28.75%) realize 2 schemata. The class of verbs having 4 to 6 schemata is fairly less numerous (443, i. e., 29.65%); those with more than 6 schemata are negligibly rare (56, i. e., 3.75%). An average verbal unit has 3.08 sentence schemata (3.13 for verbs, 2.02 for quasi- verbs).

It can be shown that the number of syntactic schemata for a given verb provides one with a good measure of the degree of ambiguity of this verb. By saying so, I do not want to claim that for each meaning of a polysemic lexical unit a special set of syntactic features exists, different from sets connected with other meanings. However, the examination of 100 units taken from the database has shown that only in 4 cases the number of meanings is higher7 than the number of sentence schemata, while the opposite holds true for a majority of entries.

Should the hypothesis of correspondence between syntactic and semantic features prove well-grounded, sentence schemata calculation could be treated as a source of data on polysemy. For the quantitativist, a syntactic approach to semantics is hard to overvalue, as it is easier to grasp the syntactic potence of a unit examined than to delve into semantics.

7. The prospects

The data can also be regarded as a starting point for further investigations. Let us mention some questions to be answered:

(a) How do syntactic properties of a given verb depend on its aspectual characteristics?
(b) Do reflexive verbs differ in a systematic way from their non-reflexive counterparts?
(c) Is the repertory of 31 sentence schemata sufficient?

There is another important problem to cope with, that of the textual frequency of sentence schemata. This problem is of a quite different character, practically and theoretically. A project has recently been spelled out at Warsaw University, aimed at syntactic analysis of large corpora of Polish texts. Its results will, I hope, be interesting both to quantitative and statistical linguists, same as those presented in this paper.

The database is written down as a dBASE III PLUS file. Each record contains 4 morphological fields, carrying - besides the headword - part-of-speech, aspectual, and conjugational information, plus 21 fields corresponding to particular sentence schemata.

In sentence schema fields grammatical characteristics of the constituent phrases are given, usually as variables. For 2-place schemata, pairs of symbols are introduced. There is, too, a numerical field which gives the number of schemata included in the entry at issue.

6. Results

The database has provided our team with a number of interesting lexicographical data. They are, first of all, complete lists of fixed values of some syntactic parameters:

(a) a list of governed prepositions in Polish,
(b) a list of types of required clauses,
(c) a list of pairs of values of various parameters in a schema,
(d) a list of idiomatic sentence schemata.

Besides these, also quantitative results are important for lexicographic purposes. The lexicographer should know which schemata are frequent in the vocabulary and which are negligibly rare - to give the latter in respective entries, while the former are present in the introduction.

The sentence schemata given in 5. vary in systemic frequency. The figures below state how many of the 1494 lexical units fulfil the various schemata. Among them, 1385 are verbs (92.70%), and 109 are quasi-verbs (7.30%):

Verbal sentence schemata:

V-0: 538 (36.01%), V-11: 1005 (67.27%), V-12: 588 (39.36%), V-13: 25 (1.67%), V-14: 9 (0.60%), V-15: 298 (19.95%), V-16: 16 (4.02%), V-17: 213 (14.26%), V-21: 412 (27.58%), V-22: 581 (38.89%), V-23: 2 (0.13%), V-24: 8 (0.54%), V-25: 3 (0.20%), V-26: 11 (0.74%), V-27: 228 (15.26%), V-28: 18 (1.20%), V-29: 92 (6.16%), V-210: 86 (5.76%), V-211: 29 (1.94%), V-212: 37 (2.48%), V-213: 143 (9.57%).

Quasi-verbal sentence schemata:

Q-0: 15 (1.00%), Q-11: 35 (2.34%), Q-12: 20 (1.34%), Q-13: 16 (1.07%), Q-14: 23 (1.54%), Q-15: 18 (1.20%), Q-21: 14 (0.94%), Q-22: 23 (1.54%), Q-23: 20 (1.34%), Q-24: 12 (0.8%), Q-25: 4 (0.27%), Q-26: 3 (0.20%).

As it is easy to see, only few of the 31 sentence schemata are frequent in the vocabulary (cf. V-0, V-11, V-12, V-21, V-22); extremely rare schemata are V-23 - V-26, to say nothing about those of the Q-type in general. Similar figures could be given for the required values of the syntactic parameters (like case, preposition, preposition-case, type of clause), as well as for pairs of the values of these parameters. Some of these values or pairs of values are required by numerous verbs, others seem more or less exceptional.

Literature

Gruszczyński, W. (1989): Fleksja rzeczowników pospolitych we współczesnej polszczyz'nie pisanej (Inflection of Polish common nouns in modern written Polish), Wrocław.
Saloni, Z. (1989): Projet d'un "Bescherelle" polonais, Revue que'becoise de linguistique, Vol. 17, No 2, 217-236.
Saloni et al (1990): Z. Saloni, S. Szpakowicz, M. Świdziński, The design of a Universal Basic Dictionary of Contemporary Polish, International Journal of Lexicography, Oxford, vol. 3, No. 1 (1-22).
Saloni, Z. and Świdziński, M. (1987): Składnia współczesnego języka polskiego (Syntax of modern Polish), Warszawa.
SJP Dor. (1958-69): Słownik języka polskiego (The dictionary of Polish), ed. by W. Doroszewski, Warszawa, vol. I - X plus Supplement.
SJP PWN (1978-81), Słownik języka polskiego (The dictionary of Polish), ed. by M. Szymczak, Warszawa, vol. I - III.
Słownik generatywny (1980-): Słownik syntaktyczno-generatywny czasowników polskich (The syntactic generative dictionary of Polish verbs), ed. by K. Polański, Wroc~aw, I (1980), II (1984), III (1988).
Słownik-minimum (1988): Słownik-minimum polsko-angielski i angielsko-polski (Polish-English and English-Polish pocket-size dictionary), ed. by K. Billip and Z. Chociłowska, Warszawa.
Szpakowicz, S. (1986): Formalny opis składniowy zdań polskich (Formal syntactic description of Polish sentences), Warszawa, 2nd ed.
Świdziński, M. (1991): Gramatyka formalna języka polskiego (Polish formal grammar), Warszawa.
Świdziński, M. and Szpakowicz, S. (1991): Sentence schemata in the Universal Basic Dictionary of Contemporary Polish, International Journal of Lexicography, vol. 4 (to appear).
Tesnie're, L. (1957): Eleme'nts de syntaxe structurelle, Paris.

Notes

1. A new description of Polish nominal and verbal inflection is proposed in Gruszczy~nski (1988), Saloni (1989). Sentence patterns for verbs are presented in Saloni and Świdziński (1987: Ch. XII) and Świdziński and Szpakowicz (1991). Other aspects of the grammatical information are still under way.
2. A syntactic-generative dictionary of Polish verbs which is now being published (Słownik (1981-)) is but an exception that confirms the rule. It is designed for highly qualified linguists, not for average users.
3. Note that so-called subordinate complex sentences are, on the basis of (ii), simple sentences.
4. There are, too, other realizations of the finite phrase. They will not be mentioned here since what I deal with here is the Polish lexicon, rather than grammar.
5. The phenomenon of quasi-verbs is Slavic-specific; nothing of this sort exists in English or German, but does exist in Spanish or Italian.
6. Note that a great Polish dictionary, edited by W. Doroszewski (SJP Dor. (1958-70)), includes 20,000 verbal entries. It should be emphasized that the selection of 1500 items at most meets the condition of structural sufficiency. Since Słownik-minimum (198x) contains, by definition, mainly most frequent Polish verbs, the statistical representativeness of this selection can be questioned.
7. Understood as a number of meanings in the respective entry of SJP PWN (1978-82).

Tests of Hypotheses Concerning Phenomena of "Computer Talk"

Extended Abstract

Dr. Christa Womser-Hacker
University of Regensburg
Dept. of Linguistic Information Science
P.O.Box 397
W-8400 Regensburg

Tel. (0941) 943-4195
Fax: (0941) 943-2305
E-Mail: womser@vax1.rz.uni-regensburg.dbp.de

0 Introduction

This paper reports on statistical hypotheses tests of simulation experiments on information seeking man-machine-dialogues in natural language, carried out at the Laboratory of the Department of Linguistic Information Science at the University of Regensburg (LIR).
On the basis of a flexible model of an information system we achieved results concerning the behavior of the human partner in man-machine-interaction. It could be shown that man-machine-dialogues differ from man-man-communication.
The DICOS project (granted by the German Ministry for Research and Technology) which started in March 1988, included two different application domains: information seeking dialogues within a railway information system and within a library environment.
This paper concentrates on the statistical part of the project. Other aspects and further details are discussed in Krause/Hitzenberger 1991.

1 Method

1.1 Experimental Design

In the DICOS project two hidden operator experiments were conducted, in each of which four information systems with different capabilities were simulated.
For each application domain, railway and library information, the four systems tested represented the capabilities of present day and future information systems.

The systems tested vary as follows:

System 1:
The user is told that he is communicating with a railway or library employee, who does the lookup in a database and answers his questions through a computer as a special input/output channel. The system has no restrictions in language utterances or cooperativity. The user gets an echo describing his information need.

System 2:
System 2 behaves exactly like System 1, but the user is told that he is communicating with a computer which understands natural language without restrictions. The user can ask what ever he wants. The only difference between System 1 and System 2 is the mental concept of the systems in the head of the user.

System 3:
The cooperativity of System 3 is restricted. The user has to formulate his information need according to given restrictions (no vague or modal expressions, literally interpretation of time phrases etc.). If the user does neglect the restrictions, his utterances are rejected and he is told by the system which mistake has occurred.

System 4:
System 4 introduces additional restrictions. It otherwise behaves like System 3, but performs no error analysis.

The experimental design involved two experimental variables or factors:
the manner of input (voice input or keyboard input) and
the system variation (System 1, System 2, System 3, System 4) as discussed above.
This led to an experimental design with eight cells and five subjects each:

	System 1	System 2	System 3	System 4
voice	5	5	5	5
keyboard	5	5	5	5

In experimental work the experimenter manipulates the experimental or independent variables in order to determine the effect on the dependent variables. What are these independent variables and how to operationalize them?

When doing the first experiment hints on the difference between manman-machine- and man-man-communication already existed. Empirical evidence on a possible language register "computer talk" has been given by Krause 1991.

In the library domain which was the second experiment the dependent variables exclusively referred to features of "computer talk". This means that all hypotheses tested statistically were based on a general assumption of the existence of a language register "computer talk". In this context, features were choosen only when there was a distinct tendency within the qualitative protocol analysis of the railway domain (cf. Kritzenberger 1990). The complex variable "computer talk" represents the sum of different single features depending on input mode and on the interpretation of the value of the different criteria.
Concerning keyboard input, the "computer talk" variable contains the following criteria:

- increase of deviant or odd looking formulations
- increase of overspecifications
- increase of deviant coordination constructions
- reduction in the usage of particles
- reduction in the usage of marks of politeness
- reduction in the usage of dialogue organizing marks
- reduction in the usage of partner oriented dialogue signals
- increase of formal symbols
- modifications concerning syntactic patterns

Examples for these criteria can be found in detail in Kritzenberger 1990.
Measurement of increase and reduction results from the compari-

son with System 1, representing the human-human-commmunication via computer (see above).

With regard to spoken input the "computer talk" variable is supplemented by typical features of spoken language (repetition of sentence fragments, different types of elliptical constructions etc.) which are interpreted as dialogue disturbances.

Statistical experiments set up hypotheses so that comparisons concerning the factor levels can be carried out by so called hypothesis tests. Within these tests the null hypothesis which says that there is no difference between the factor levels is compared with another hypothesis. The null hypothesis can be rejected, if a difference has been shown within a specified probability of making a wrong decision.

According to the variables shown above statistical hypotheses were formulated which are to contrast a human-human-communication situation with a human-computer dialogue situation (S1 vs. S2, S2, S3), a cooperative human dialogue partner with a cooperative computer system (S1 vs. S2), a restricted communication situation with a non-restricted situation (S1, S2 vs. S3, S4) and voice input with keyboard input (S1, S2, S3, S4 voice vs. S1, S2, S3, S4 keyboard), e.g.

$$H1: \mu_{CTA}(S2,S3,S4;\ \text{keyboard}) > \mu_{CTA}(S1;\ \text{keyboard})$$

H1 postulates that the total number of "computer talk" features is lower in communication with a human than with a computer system.
This hypothesis contrasts system 1 with the other systems tested.

1.2 Realisation

For each experiment (railway and library) 40 test persons were recruited from students of the university. They had to perform eight tasks, which took them about two hours. The choosen tasks represented different scenarios as shown in the example below:

> "Your infirm grandmother wants to travel to Hamburg. Find a travelling possibility as convenient as possible for her".

According to the experimental design, 20 test persons had to perform the tasks using keyboard input, the others using voice input. In order to be able to count the different features, the sessions were recorded.

1.3 Statistical tests

To show the effect of the two factors (input and system) on the dependent variables a multifactor analysis of variance (ANOVA) was performed. ANOVA presents several sums of squares which measure the amount of variation in the data. Within a specific significance level, ANOVA analyses the effect of one or more factors on the response variables.
ANOVA makes clear whether there is a significance due to the two factors, to the factor interaction or to the residual.

To show differences between the factor levels (the cells) the Scheffé test was applied.

2 Results

Our hypothesis 1 that the total number of those features of "computer talk" whose values increase when talking to a computer is higher in System 2, System 3 and System 4 than in System 1, could not be proved. This is valid for both input modes.

The features whose reduction signals "computer talk", showed significant results concerning voice and keyboard input. There is a significant difference between man-machine and man-man-communication in the usage of these features.

The analysis of variance also showed significant results with regard to the partner oriented dialogue signals. People use more of these signals when talking to a human partner than when they talk to a computer system which behaves exactly like the human informant.

According to statistics no significant difference could be proved with regard to the usage of the formal symbols and the variation of the syntactic patterns.

3 Conclusion

The experiments conducted in the DICOS project led to a statistically plausible result. On the basis of the influence of the language restrictions on the one hand and of the user´s mental model of his dialogue partner on the other, the existence of some "computer talk" features were proved.

Literature

Krause, J., Hitzenberger, L. (Hrsg.) (1991), Computer Talk, to appear.

Krause, J. (1991), Empirical Indications about the Existence of a Language Register "Computer Talk". In: Bammesberger, A., Kirschner, T. (Hrsg.) (1991), Language and Civilization. A Groudwork of Essays and Studies in Honour of Otto Hietsch, to appear.

Kritzenberger, H. (1990), Zur Simulation informationsabfragender Mensch-Maschine-Dialoge in natürlicher Sprache Im Projekt DICOS. In: Herget, J., Kuhlen, R. (Hrsg.) (1990), Pragmatische Aspekte beim Entwurf und Betrieb von Informationssystemen. Proceedings des 1. Internationalen Symposiums für Informationswissenschaft, S.172-183.