QUANTITATIVE LINGUISTICS

Vol. 49

Glottometrika 13

Burghard Rieger (Hrsg.)

Universitätsverlag Dr. N. Brockmeyer
Bochum 1992

QUANTITATIVE LINGUISTICS

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Glottometrika ... – Bochum : Brockmeyer.
Früher mehrbd. begrenztes Werk
ISSN 0932-7991
13 (1992)
(Quantitative linguistics ; Vol. 49)
ISBN 3-8196-0036-1
NE: GT

ISBN 3-8196-0036-1

Uni-Tech-Center, Gebäude MC, 4630 Bochum 1
Gesamtherstellung: Druck Thiebes GmbH & Co. KG Hagen

Contents

Rieger, Burghard (ed.)
Glottometrika 13
Bochum: Brockmeyer 1992, 1-104

Metalinguistic Integration under Conditions of Uncertainty: Second Language, Metaphor and Idiom Comprehension

Joel Walters and Yuval Wolf

Contents

PART I: THEORETICAL BACKGROUND

CHAPTER ONE: Introduction

1.1 Background and Overview

This volume is an attempt to provide the reader with a description of some empirical work conducted to explore integrative aspects of the process of second language use. Our work on psycholinguistic information integration does not, at present, afford a completely comprehensive view of the field. We heuristically sampled different aspects of second language (SL) use, such as integration of first and second language information in lexical comprehension, integration of information about topic and vehicle in comprehension of code-switched metaphors, and comprehension of unfamiliar idioms in a second language. The disparity of topics can be related to the heuristic and exploratory nature of the research conducted. In that vein, we had a notion that the field of SL use might benefit from an approach conducted in terms of information integration. We conceptualize SL use as a process which entails integration of different sources of information, both textual and extra-textual. We view this process as an attempt to comprehend relatively unfamiliar material on the basis of missing information. Fortunately, five out of six of our explorations appear to be reportable. (The sixth dealt with comprehension of fables, and our failure there was primarily due to technical details).

In Part I we review several current models of second language use and attempt to examine their connection to our interest in integrative processing as a conceptual framework for the study of SL comprehension. Part II reports on empirical work illustrating this approach. Part III elaborates on several substantive and methodological topics arising from our conceptualization of SL use as an integrative process and its experimental treatment in the studies presented.

1.2 Definitional Preliminaries

At the outset it might prove helpful to clarify the focus of this research along a number of dimensions. First, reading comprehension is seen here as a part of the domain of language use and only tangentially related to language acquisition. Comprehension does rely on phonetic, syntactic, and semantic properties of

language; its focus, however, is pragmatic. In other words, the impetus for comprehension is the desire for understanding and communication — a pragmatic process. In a way, comprehension, when viewed from the perspective of language use, can be seen as a prerequisite or precursor to the study of language acqusition or learning. The comprehension of a lexical item, syntactic structure, or pragmatic expression can be viewed as a single event, while language learning constitutes a series or composite of events. In linguistics, then, comprehension fits properly into the study of pragmatics.

In psychological terms, we adopt Anderson's (1981:212) "functional view of language in which meaning lies in the communicator rather than in the communication." In this view, the language user "attempts to convey a meaning or referent, and tries to delimit it by using a number of words. The intended meaning is not the referent of any one word, but of all the words, not to mention associated non-verbal cues and context-background information." We can infer from this that the listener or comprehender of language is motivated to communicate fully with the speaker, and thus attempts to integrate information from various relevant sources.

With regard to the relevance of second language (SL) use to bilingualism, we follow Klein (1986) and others (Lamendella, 1977; Rutherford, ms; Sharwood Smith, 1986) in opting for a distinction between primary (L1) and secondary (L2) language use. While most of the data cited are taken from subjects which have traditionally been labeled FL users (due to the social context in which their second language is used) the FL/SL distinction is deemed unnecessary here for two reasons. From one point of view, FL comprehension is viewed as the simplest case of SL use (cf., Klein, 1986:19, who states that "SL also covers FL"), lending itself to be studied without the heavy influence of social, cultural and cognitive factors so crucial in SL studies.

More importantly, our interest in comprehension mitigates the need for any distinction between FL and SL use. Comprehension is viewed here as a psychological event, observable at a given point in time. As such, FL and SL are subject to the same constraints and processes. Thus, while the focus here is for the most part on FL comprehension, no a priori theoretical distinctions between FL and SL comprehension are made. Moreover, there are no claims about the generalizability of our findings to SL use. By the same token, the intent here is not to say anything about SL production, and the studies of comprehension reported here

are only a tentative beginning to a program of research on various aspects of integration in second language reading comprehension. For convenience, we adopt the use of the term SL to refer to secondary or foreign language and PL to apply to primary, first, or native language use.

CHAPTER TWO: Theoretical Approaches

This chapter is divided into five sections. At the outset we review Weinreich's classic work, focusing particularly on the aspect of this work which relates to integrative SL processing. Next contrastive analysis and its offspring, error analysis, is considered. The following two sections present several current approaches to the study of SL use which bear upon questions of interest in the empirical work reported on here. Finally, the most recent work in the SL field, conducted in a universal grammar framework, returns us to the basic issues of contrastive analysis.

2.1 Weinreich on Integration

The present reexamination of Weinreich's work considers his views on the role of PL and SL structures in the description of bilingual speech, or what Weinreich himself calls 'linguistic integration.' Following a declaration of interference as his primary linguistic interest, Weinreich states that this concept "implies the rearrangement of patterns that result from the introduction of foreign elements into more highly structured domains of language" (1967:1). He cites Vogt (1949) who writes in the same vein of a "reorganization ...of the system."

Weinreich (1967:7) distinguishes between two types of interference. One is borrowing or transfer, which involves categorization of an utterance as belonging to a particular language and a clear distinction that it contains foreign elements. The second type of interference involves what Weinreich describes as "interlingual identifications," which result in an overlapping of two linguistic systems. He provides phonemic, syntactic, and semantic examples of this phenomenon, concluding with two hypotheses parallel to the two types of interference. These approaches are labeled "two coexistent systems" vs. "a merged single system." Weinreich himself tends to favor the "separate systems" analysis to describe the speech of bilingual and a single system for describing the language itself, at least at the phonemic level (1967:9, footnote 6).

These two competing approaches also find expression at the lexical level. Based on the distinction between speech and language, which is so essential to the Saussurian tradition, and more specifically the distinction between signified (content) and signifier (expression), Weinreich describes three types of bilingualism: coordinative, compound or merged, and subordinative. In the first type, a

word or sign has two coexisting signifiers, one for each signified. In compound bilingualism, there is one signified with two ways to express it. In subordinate bilingualism there are also two forms to express a single content, this time the expression in L2 matching an already known form in L1 and its corresponding content. This taxonomy affords a basis for predicting differences in how new words are learned as well as how they may be stored and/or retrieved. The first two approaches mentioned represent the two competing postitions which have served as the basis for much of research in the psychology of bilingualism.

Terminological changes have been one of the hallmarks of this tradition. Weinreich himself seems to have preferred the distinction between coexistence and merging. Ervin and Osgood (1954) and Lambert (1961) were among the first to investigate this distinction experimentally and did so under the terms coordinate and compound bilingualism. Later, as the focus of psychological research became memory and its organization, bilingual research followed suit.

The two principal competing positions parallel our interest in psycholinguistic integrative processing, the coordinate/separate/ independent/dual system representing a non-integrative hypothesis and the compound/merged/interdependent/single system corresponding to the integrative approach. The questions which emerge from both linguistic and psychological approaches, including whether the language user has one or two linguistic systems, how the lexicon is organized for bilinguals, how new words are integrated into that lexicon, etc. seem to reflect the same basic interest which Weinreich pioneered.

2.2 First Generation Models

Contrastive Models

A number of overviews of the SL field (Brown, 1980, 1987; Dulay, Burt, and Krashen, 1982) indicate that contrastive analysis developed out of the tradition of structural linguistics. Fries (1945), one of the proponents of this latter tradition, described second language learning as being quite different from first language learning. He argued that difficulties in SL learning were not identifiable through analysis of SL production. Rather, they could be traced to a habitual set created by PL patterns. Following in this tradition, Lado (1957) maintained that, through a systematic comparison of the native language and culture with the target language/culture, these patterns of difficulty can be both described and predicted. He also claimed that those patterns which would not present any difficulty to SL learners could be predicted in a similar manner.

With the rise of Transformational-Generative Grammar, there were several attempts (Ritchie, 1967; Wardaugh, 1970; Wolfe, 1967) to dismiss the contrastive analysis (CA) approach. Paradoxically, research on CA was revived through one of these attempts, Wardaugh (1970) distinguishing between a 'weak version' of CA and a 'strong version' of the theory. The former maintained that a comparison of first and second language structures could describe difficulties in SL learning. The latter claimed that, through a structural comparison of PL and SL, one could predict a learner's errors.

The heavy demands made by CA on linguistic analysis led to certain problems in the research program. The hypothesis derived from CA states that a complete description of both PL and SL, whatever those languages might be, is possible, or that at least a complete description of learner difficulties can be undertaken. CA also assumes that such a description and comparison of PL and SL will predict degree of difficulty in learning. Furthermore, CA presupposes an operative measure for learner difficulty which is both valid and reliable (Fraser, ms).

CA's connections with structural linguistics brought about criticism as linguistics became more mentalistic and generative. That criticism (Brown, 1980; Dulay, Burt and Krashen, 1982; Ritchie, 1967; Wardaugh, 1970) questions CA for its fundamental assumption that new habits are what account for SL learning. These habits include new (SL) phonological, morphological, syntactic, and semantic patterns as well as new lexical forms. The criticism also reflects certain difficulties with concepts such as transfer and interference. Early CA thinking described positive transfer between two languages as a sharing of habits. Negative transfer was said to be the result of linguistic interference.

The terminology here does not make an explicit distinction between linguistic and psychological contributions to the problem. The controversy over CA may come in part from a lack of clear boundaries between its linguistic and psychological claims. In other words, both disciplines have played a role in the development of CA's research program. These two disciplines do not always address the same issues. Nor are their conceptual and operational terminologies necessarily the same.

The controversy between behaviorism and nativism surrounding CA may be more polemics than an expression of fundamental differences. Carroll (1968) expressed this view in a review of the field of CA. There he points to similarities between nativism's rules and rule-governed performance, on the one hand, and behaviorism's habits and responses, on the other. Lado's own work (see, for example, Lado, 1968) develops from reliance on a structuralist/behavorist

tradition to adoption of "Mentalistic Theory" as one of its bases. Finally, James (1980) points out that the parallels between the two schools may be more numerous than the divergences. He illustrates this by detailing the similarities between transfer theory and Gestalt psychology's notion of 'set.' All of the above suggests that the behaviorist/nativist controversy may no longer be a live issue, at least not with regard to the viability of a contrastive approach to SL use.

While much of the controversy over CA has focussed on terminology, Carroll (1968) and James (1980) treat the role of transfer and interference in SL learning. Carroll comes to the conclusion that research which has been conducted under the heading of transfer or interference theory can contribute relatively little to the field. He reports that most experiments examining transfer deal with a new or different set of habits that replaces previous patterns. In contrast, he argues that second language use (as compared to non-linguistic transfer phenomena) involves "transfer from a very highly learned system of habits to a new and different set" (Carroll, 1968:117). The implication is that both PL and SL skills are involved in second language performance despite the apparent unidirectional tone (i.e., from PL to SL).

James (1980), in advocating contrastive analysis, presents transfer and S-R theory as the psychological bases for that approach. With regard to the former, he deals with both transfer of PL features to the SL as well as with what Jakobovitz (1970) terms the "backlash" of SL on PL performance. In discussing S-R theory, James prefers 'mental conditions' to Jakobovitz's more behaviorist orientation, where 'environmental conditions' are prior to utterances as descriptive of the stimulus information.

In considering the phenomenon of transfer, James cites the work of Corder which, like that of Lado above, remained substantively the same, although it underwent superficial terminological changes. In 1971 Corder describes the learner's behavior as influenced by mother tongue habits, while in 1975 he calls the same phenomenon cognitive structures.

Error Analysis

Corder's (1967) paper is commonly accepted as the beginning of the field of Error Analysis (EA). This paper attempted to show that SL errors came from more than a single source (i.e., only PL). The underlying (and sometimes even explicit) assumptions of the field were that errors in SL speech were inevitable, but that with enough effort on the part of the learner could be eliminated.

Among the types of errors classified in the EA research paradigm were: interference errors (i.e., those which could be predicted or identified through a comparison of first and second languages), developmental errors (i.e., those which could be predicted or identified through an analysis of the regular course of child language development), overgeneralization, hypercorrection, spelling pronunciations, cognate pronunciations, frozen forms, analogies, and simplifications (see Tarone, Cohen, and Dumas, 1976, for a review).

The primary focus of the field was on the second language user's speech, i.e., on production data. There was, however, some concern with the origins of errors. McLaughlin (1987:92-94) reviews these causes in the context of a discussion of interference. His taxonomy includes: dominance of one language, mixing, similarity between L1 (PL) and L2 (SL), lack of differentiation among domains of use, minimal contact with native speakers, and foreignness of the social environment. While some of these refer explicitly to linguistic structure and others to sociolinguistic phenomena, the notions of dominance and mixing have a psycholinguistic tone, indicating an interest beyond the simple product of SL and towards more processing oriented terminology.

Interlanguage

The EA research program proliferated rapidly into a series of related subfields with an overlapping interest which came to be called interlanguage studies, each maintaining its own slightly different perspective. The studies in this domain included the search for and classification of learning strategies, communication strategies and cognitive strategies in SL use. The focus here went beyond the product of SL processing, although production data was still the primary source of inference for identification of strategies.

Among the learning strategies discussed in the literature are: transfer from L1 (PL) unless proven wrong (Corder, 1973), avoid hard things (Schacter, 1974), memorize, imitate, be fluent, seek reinforcement, maximize functional load (Fillmore, 1979), and maximize input. The orientation of these strategies is focused on the SL learner (see Brown, Crymes, and Yorio, 1977; Dulay, Burt, and Krashen, 1982; and Oller and Richards, 1973 for a representative sample of the work current at the time). Like the latter attempts, identification and investigation of communication strategies derived from a learner's perspective, reflecting some of the same interests as those just described.

The communication strategies most frequently referred to were: paraphrase, avoidance, simplification, transfer, overelaboration, prefabricated patterns, ap-

peal to authority, and language switch (Tarone, Cohen, and Dumas, 1976). Finally, in a further attempt to organize and synthesize SL production data, cognitive strategies brought the center of attention, already on the learner, a more psycholinguistic perspective. These strategies, similar in name as well as in intent to those listed in connection with the other approaches include: overgeneralize target language rules and semantic features, use context (i.e., assume the conversation is relevant to the situation), use formulaic speech, search for recurring parts, be fluent, and work on big things and save the details for later (Fillmore, 1979).

Error analysis and interlanguage studies succeeded in overthrowing CA as the major research trend in the study of SL use. In doing so, these approaches opted for a more abstract terminology (see Tarone, Cohen and Dumas, 1976). The abstractness of that terminology raised questions about whether process or product was the main concern of the SL field. In retrospect, it is fairly evident that the products of SL use, i.e. the actual utterances spoken, were the primary focus of error analysis, while learner strategies, as inferred from SL production data, became the chief concern of interlanguage studies.

Despite these differences in approach, we see some overlap in the substantive issues addressed in the period from the rise of CA through error analysis and interlanguage. All showed concern for the role of a PL in second language use. Some labeled this concern transfer (CA, learning strategies, communication strategies); others called it interference (CA, error analysis). Similarly, the differences notwithstanding, the fundamental grounding in descriptive linguistics is a common feature of CA, error analysis, and early interlanguage studies alike.

2.3 From First to Second Generation Models

Most current theories in second language acquisition do not share all of the same issues with researchers in second language use/comprehension. Several models, however, stand out as exceptions. In particular, one model which considers the role of PL knowledge, presenting a sort of interactive approach between that knowledge and SL information, is considered. That model (Klein, 1986) lays out a theory of analysis and synthesis based on knowledge available to the learner and "structural properties of input" (p.63). Klein's proposal makes explicit two features of the empirical work conducted and report on below. One is the role of PL knowledge in SL use, and the other is the interactive nature of the process.

Klein's (1986) classification of available knowledge includes the following four sources: general linguistic, non-linguistic or contextual, SL, and PL structural knowledge. The latter source of information mentioned, PL knowledge, is discussed in traditional Contrastive Analysis terms: "The learner's knowledge of a first language may have positive as well as negative consequences for the problem of analysis" (Klein, 1986:64).

This formulation reflects constructs such as positive and negative transfer (interference) which are characterizing features of CA. In his disucssion of CA, Klein dismisses the so-called Strong Version of the theory. He concludes, however, that the use of PL knowledge in SL use is undeniable. In this context he states: "Whenever a learner of a second language tries to comprehend or to produce utterances in that language, he relies on all sorts of knowledge that might help him. One component of this knowledge is what he knows about his first language..." (Klein, 1986:27).

The interactive nature of SLA, as Klein describes it, entails a functional view of linguistic knowledge. By functional he means that the different sources of knowledge are inter-related, interdependent and synchronized (1986:48). Two of his six dimensions of language acquisition are particularly relevant here. One is what he calls the language faculty or language processor, which is among the three dimensions which 'determine' the language acquisition process. The other is the structure of the process. One of the unique capacities of the language faculty Klein mentions is the potential for "adjusting its language production and comprehension to the particular linguistic material" (1986:39). With regard to SL acquisition, this unique capacity finds expression in the potential "to reorganize [Klein's emphasis] the language processor to cope with another language..." (1986:39)

In the structural dimension this reorganization ability takes the form of 'synchronization' between and among various sources of knowledge. This synchronization or interaction is said to operate on both the external and internal levels. On the external level, Klein specifies the interaction of linguistic and non-linguistic knowledge. Internally, he mentions the synchronization of phonological, morphological, syntactic, and lexical knowledge.

Thus, different sources of knowledge (general linguistic, non-linguistic, SL, and PL) are shown to operate as well as interact with structural properties of the input in the learner's analysis. In synthesis, Klein concentrates more on the interaction among the various sources of knowledge, coming to the conclusion that SLA is

a developmental process where learner varieties are eventually replaced by target language rules.

This competitive view of PL and SL knowledge may mask an intrinsically positive contribution of both PL and SL information and even the possibility of integration of these sources in SL use. The developmental/acquisitional questions notwithstanding, Klein's approach clearly admits the role of both PL and SL knowledge and certainly leads to the possibility that they are integrated. These claims are made for SLA and are based on production data in Klein's work. What holds for SL comprehension is an empirical question to be taken up later in this work.

2.4 Second Generation Models

This section deals with questions about the nature of models and theories which have been put forward to describe or explain SL processing. Krashen's (1982) Monitor Hypothesis, McLaughlin's cognitive theory and information processing approach (McLaughlin, 1987; McLaughlin, Rossman, and McLeod, 1983) and Bialystok's (1981, 1984) model of analysis of knowledge and control of cognition are selected to represent the generation of thinking which followed CA, error analysis, and interlanguage studies. All make an attempt to go beyond the level of product in their proposals and investigations of SL use. The discussion here focuses on the issues these theories address as well as their methods.

Krashen's Model of Acquisition and Learning

The most detailed descriptions to date of the Monitor Model (Dulay, Burt and Krashen, 1982; Krashen, 1982) present a three-part model involving a filter, an organizer, and the monitor itself. The organizer and monitor are a distinction between acquisition and learning, a distinction which represents the fundamental purpose of the model. In this model, as well as in an earlier discussion of it (Krashen, 1977), one way the author distinguishes between acquisition and learning is by specifying two processing differences, time and attention. The model dictates that the amount of processing time available determines which mode, acquisition or learning, the SL user will utilize. Acquisition mode is said to be employed when limited time is available for processing, while the learning mode is invoked when the SL user has more time available, and the learner uses what Krashen calls a "conscious grammar."

The other processing dimension of the model, attention, is said to be necessary in order to initiate operation of the monitor. Still earlier discussions of the model (Krashen, 1976, 1977; Krashen and Pon, 1975) express processing time as the trigger for the monitor, while Krashen (1982), in a more recent presentation of the Monitor Model, lists time, attention, and rule knowledge as the conditions necessary to activate the monitor.

In this later version (Krashen, 1982:15-20), the processing aspects of acquisition and learning are made much more explicit. Acquisition processes are claimed to be responsible for initiating utterances in a second language and for controlling fluency. Learning is restricted to one function only: to activate an editor (i.e., monitor), modifying an utterance after it "has been produced by the acquired system." (p. 15)

Among the processes mentioned above, initiating and fluency seem to be a complex of operations, which involve planning, selecting, retrieving, and ordering, among others. The Monitor Model does not capture all of this complexity, and despite the apparent attention paid to processing in this terminology, the primary motivation for this model is to provide an explanation for functioning in L2 in the context of previous experience. In the most general sense, this interest can be looked at as integration of L1 information (previous experience) with new material in L2 in the comprehension and production of a second language.

With regard to our interest in a potential role for PL in SL processing, Krashen (1981) reviews data from Duskova (1969), LoCoco (1975), and his own work, concluding with the following generalizations:
1. L1 influences the acquisition of word order;
2. L1 shows weak influence in the area of bound morphology;
3. L1 influence is strong in 'acquisition-poor' environments (e.g., among adults as opposed to children and in foreign language as opposed to SL contexts).
Krashen sees PL knowledge as a 'fall-back' option for the SL user in production when faced with limited SL knowledge and skills. Since acquisition is viewed as an ideal, and the PL is claimed to play a smaller role in 'acquisition-rich' environments, the author proposes reduction or elimination of PL influence as unnatural. Thus the model reviewed here leaves no room for the empirical question raised here with regard to integration of PL and SL information in SL processing.

From the point of view of methodology, Krashen's model raises a question about the individual differences/universals dichotomy. The Monitor Model has drawn almost exclusively on data from elicitation of grammatical morphemes (Bailey,

Madden, and Krashen, 1974; Dulay and Burt, 1974; Larsen-Freeman, 1975; Krashen, 1977) and individual case studies (Krashen, 1978; Krashen and Pon, 1975) for evaluating use of a monitor, or conscious grammar. The problems with these techniques have been discussed at length elsewhere (e.g., Larsen-Freeman, 1976; Rosansky, 1976) and include the learning environments sampled, the elicitation procedures employed, and the scoring methods used.

The 1982 presentation of the model places heavy stress on individual and group differences. In that book, Krashen distinguishes among over-users, under-users, and optimal users of "conscious rules." His ultimate conclusion is that only acquisition mode can promote progress in SL. This conclusion also implies that the PL is irrelevant to second language use and that only pure experience with the SL can bring about real performance gains. The distinction among monitor users carries with it the attitude that under-users of the monitor make greater progress in SL performance than over-users and that optimal users make still greater progress than these other two types.

A critical look at the user types inferred from the kinds of data mentioned above (i.e., morpheme analyses and individual case studies) and other such post-facto treatments are not likely to be able to provide answers to questions about the relationship of monitor use to individual differences. More specifically, the data cannot determine whether an individual is an overuser or an underuser or whether other situational/processing differences (i.e., overuse or underuse at a given time or under any particular condition) can provide an adequate account of the data. This raises questions about the use of data from individual differences in the study of SL use. Specifically, one of these questions is whether distinctions between over-users and under-users are adequate enough to explain, or even describe, how a second language user attends, plans, selects, retrieves, identifies, discriminates, classifies, comprehends, or understands verbal material in a foreign language?

Nevertheless, our interest remains ultimately in the behavior and processes exhibited by the individual language user. In this light, caution is called for with regard to assumptions on which the research is based. Minimally, this includes matched samples, reliable administration, and consistency of measurement. This is not to imply that the search for individual differences in SL processing lacks methodological rigor. Rather, prudence in the study design as well as in the interpretation of the data is what seems to be needed. In particular, any research question should be divisible into its major components, including its time units, its relevant variables, and its measures, among others. Thus, in order to provide

valid answers to questions about SL use, an individual differences approach might profit from complementary treatment with a rigorous experimental analysis.

McLaughlin's Information Processing Approach to SL Use

McLaughlin (1987) treats three concepts of interest here: transfer, automaticity, and restructuring. The first is discussed in the context of his treatment of interlanguage theory, the latter two in a presentation of the contribution of cognitive theory to second language learning. While McLaughlin's focus on learning is not our topic here, the concept of transfer is of interest for its potentially interactive contribution to SL use, while automaticity and restructuring would appear to play a role in processing of all types, comprehension and thinking not the least among them.

The concept of transfer is presented by McLaughlin to show the importance of a first language in SL learning. In contrast to much work on interlanguage as well as the claims of Krashen, Dulay and Burt (1982), which tended to ignore or play down the role of the PL, McLaughlin, like Klein (1986) and others (see, for example, Zobl, 1983, and the work on markedness and universals from Hyltenstam, 1982, and Kellerman, 1983) indicate a revival of CA as a potential contributor to an understanding of SL use.

McLaughlin examines transfer as a process which takes the notion beyond its original conception as simply the product of SL speech or writing. He reviews evidence of studies of similar developmental sequences, but differences in rate of acquisition (Keller-Cohen, 1979), qualitative differences in the acquisition process exhibited by various PL speakers learning the same SL (Zobl, 1982), and differences in fossilization as a function of whether certain structures can be eliminated from SL speech. In spite of the overt interest in process, all of these data come from an examination of SL production, process being inferred from product. Nevertheless, the evidence is commanding with respect to the role of the PL in second language use, thus justifying the revived interest in CA.

In a more overtly process-oriented statement, McLaughlin also presents transfer as a type of decision-making in which two factors, cited from the work of Kellerman (1979, 1983), are involved. One is the "perception of the similarity between first- and second-language structures" (McLaughlin 1987:79); the other is the extent of markedness of the PL form. Both are said to be involved in a decision-making process which determines whether or not a structure is transferred from the PL to the SL. Although the verbal material which serves as a

basis for this decision is a product of language use, the decision itself seems to be a complex process, involving information from two sources (similarity and markedness).

In his evaluation of this work, McLaughlin refers to yet another kind of interaction, i.e. the interaction of various sorts of linguistic information: phonological, morphological, etc. This innovative representation of the transfer concept leaves us with two ideas of major import in SL use: that the PL cannot be dismissed out of hand as a source of information in SL use and that SL processes seem to involve interaction of information from several sources.

The two other issues of interest here, automaticity and restructuring, are presented in the context of an information processing approach to second language learning. Again, our focus here is not on learning. Nevertheless, some of the same processes discussed by McLaughlin would appear to be involved in SL comprehension. Automaticity is defined by McLaughlin, following Shiffrin and Schneider (1977), as "the activation of certain nodes in memory every time the appropriate inputs are present" (1987:134). It is described as facilitating skill integration, an operation necessary because of the limited processing capacity of the human organism. He reviews evidence from studies in bilingualism as well as second language use in three areas: lexical processing, syntactic recognition, and reading. These data are said to favor gradual development along a controlled to automatic continuum, giving way to restructuring in later stages of SLA.

Perhaps the closest concept in the SL literature to our interest in integration, restructuring is described as a means for "interpreting new information and for imposing a new organization on information already stored" (McLaughlin, 1987:136). An important element of the restructuring process, according to McLaughlin's interpretation of the work of Karmiloff-Smith (1986), is that any new organization involves a "unified representational framework." That framework is said to provide the link between old and new information, or between what Karmiloff-Smith describes as "isolated procedures" and a single system. Generally, then, it can be inferred that restructuring may play a meaningful role in the integratation of SL with PL information.

As with automaticity, McLaughlin again reviews three bodies of literature to present evidence for his notion of restructuring in SL learning: developmental patterns, learner strategies, and reading. In his own research on reading, restructuring is used as an explanation for the inability of advanced ESL learners to make full use of syntactic and semantic information in an oral reading task. Cziko

(1978), in a study of syntactic, semantic, and discourse constraints in SL reading, reports on a similar finding with regard to proficiency differences. These results lead one to conclude that even a non-fluent SL reader can go beyond the graphic information in reading.

In an earlier proposal, McLaughlin, Rossman and McLeod (1983) distinguish between individual and situational issues in SL use. Their approach here, as above, distinguishes among individual SL users vis-a-vis the degree and nature of their learning experience. The authors outline a model with two factors, one which differentiates between second language learners who are new at a skill and those who are well-trained (controlled vs. automatic) and another which distinguishes between those who base their performance on formal rule learning and those who use implicit analogic learning. While this latter distinction may not be unrelated to Krashen's two kinds of monitor users, the model does involve aspects of cognitive psychology beyond those in the domain of learning in general and in Krashen's work in particular.

Two types of processes are defined in the above model: focus of attention, described as 'focal' or 'peripheral,' and information processing ability, labelled controlled and automatic. The distinction between focal and peripheral focus of attention is further elaborated on as a distinction between intentional and incidental performance. This model draws on psychological terminologies such as focus of attention. The information processing perspective finds expression in the controlled/automatic distinction, which in McLaughlin et al.'s model is collapsed into the distinction between new and well-trained skills. Thus, the distinction between controlled and automatic information processing has no operational means of expression in the model.

Bialystok's Analysis of Knowledge and Control of Cognition

Bialystok's work shares a common interest with that of McLaughlin, both from a conceptual perspective as well as a methodological one. Her model attempts to account for variability in SL production and gradual mastery or control of the target language. It has been modified several times since the first proposal (see Bialystok 1978, 1981, 1984). The model reflects a multidimensional approach, recognizes the complexity of language processing, and attempts to combine two aspects of that complexity in a unified experimental framework. The present section considers the objectives of this model as well as the experimental tasks which have been employed in testing it.

Bialystok's (1981) model focuses on two aspects of knowledge: its analysis and access to it. In a more recent version of the research program, Bialystok (1984) embeds both analysis of knowledge and access to information in a single framework labelled metalinguistic awareness. Analysis of knowledge is defined as a skill which allows the language user to converts implicit information into explicit, usable material. The following are provided as examples of this kind of knowledge: awareness of the relationship between structure and meaning, knowledge of the unity of speech (words, syllables, phonemes), and knowledge of syntax itself. Bialystok's second dimension is called access to knowledge, or control of cognition, and involves processes such as attention and switching. In a further and more elaborate explication of the theory behind the model, Bialystok (1986) describes this component, i.e., access to knowledge, as a scale from controlled to automatic.

As in the case of Krashen's and McLaughlin et al.'s approaches, the primary focus here centers on acquisition or learning. One component (analysis of knowledge) refers to the level of skill a second language user has already acquired. The second, access to knowledge or control of cognition, treats storage and retrieval. There does not seem to be any indication that information processing may be of central importance to the model. However, this is not necessarily a shortcoming, especially since its objectives are stated to be other than those in the field of information processing. Thus, as a model which attempts to address issues regarding language mastery, the experimental tasks are most appropriate, including: (1)grammaticality judgments, (2)identification of grammatical deviance, (3)correction of syntactic errors, (4)judgments of meaningfulness, (5)awareness of words, and (6)solution of sentence anagrams (Bialsytok, 1981, 1984, 1986). The first four examine the processes involved in analysis of knowledge; the latter two treat questions of information access. All are concerned with second language learning.

2.5 First Generation Models Revisited

A full-fledged review of the flurry of activity surrounding Universal Grammar (UG) in the past five years is beyond the scope of this chapter. However, two questions raised by that work are relevant to our interest here in SL use. One is a concern with the role, if any, of PL information in SL processing. This issue is approached, sometimes directly and sometimes indirectly, in notions of transfer, universals, markedness, core vs. peripheral grammar, and parameter setting. A second question, which impinges on our interest in integration, is more psychological in nature. It is related on the one hand to the work of Clahsen (1987) and Kail (1987) on autonomous linguistic processing, and on the other hand to studies conducted within the framework of Bates and MacWhinney's (1981, 1982, 1987) competition model (e.g., Gass, 1987; Kilborn and Cooreman, 1987; McDonald, 1986, 1987). These are reviewed in Chapter 3.

Language Transfer and Universals

Kellerman (1984) presents perhaps the most thorough review of evidence favoring a role for t/he PL in SL use. His data include syntactic topics, such as negation, word order, subject pronoun deletion, preposition stranding, hypothetical conditionals, and pronominal reflexes in the use of relative clauses; discourse features such as topicalization; and issues in lexis and semantics, including loan translations/extensions, language switching, and the influence of L1 on preposition use. Kellerman also cites the work of Schumann (1981) and Meisel (1980) who document different percentages of errors in an SL corpus as evidence for and against PL transfer.

In addition to the empirical contributions of Kellerman's (1977, 1979, 1983, 1984) work on interlanguage to the notion that PLs play a meaningful role in SL performance, his critical analysis of terminology and methodology are no less important. On the former issue, he points out that the term transfer does not capture the full range of phenomena involving PL influence on SL use. He cites 'L1-based constraints' on SL forms, 'L1 as a catalyst or inhibitor,' 'L2 influence on L1,' and avoidance as example phenomena. Kellerman (1984) rejects transfer as a descriptor of these constraints, strategies and processes, opting for 'cross-linguistic influence' as more accurate.

With regard to methodology, Kellerman laments the shortcomings of many interlanguage studies, pointing out their variability and lack of definitive results. He opts for experimental techniques, citing translation, sorting, focused judgments, and paired comparisons as appropriate techniques for the study of semantic topics in transfer.

Pfaff (1984) investigated the use of the copula, perfect tense, auxiliaries, articles, and semantic extensions in the German of Turkish and Greek children. She states that "L1 transfer does not appear to play a large role." However, she goes on to describe how more subtle differences between Turkish and Greek speakers can be identified by the influence of L1. Zobl's (1984) position contrasts with that of Kellerman and Pfaff. He maintains that PLs play "a rather restricted role" in interlanguage.

McLaughlin's (1987) review of SL learning theories includes a chapter on universals in which the author contrasts a language transfer approach with two separate approaches to the study of universals: Greenberg's (1966) typological universals and Chomsky's universal grammar. Citing data from Gass (1979) and Hyltenstam (1983) on relative clauses and pronominal reflexes from a typological universals orientation and from White (1983) in the framework of universal grammar, McLaughlin concludes that both universal and language specific processes can operate simultaneously, or at least interactively. Interlanguage, then, in McLaughlin's terms, is a product of what he calls "multiple causation" (1987:88).

The studies mentioned here, all conducted in a transitionary period and within a general framework of interlanguage studies, have led the way to a more theoretically motivated approach to SL acquisition. These approaches have drawn on Chomsky's notions of markedness, core grammar, and parameters in order to search out ways for predicting learner difficulty in SL production.

Markedness and the Core/Periphery Distinction

Eckman's (1977, 1985) markedness differential hypothesis predicts difficulty in learning a SL if the PL is structurally different and if the target, or SL, is more marked. Thus, according to Eckman, an English speaker learning German is predicted to have more difficulty learning English voicing contrasts than a German speaker learning English, since English is presumably more marked on this feature than German.

McLaughlin (1987) treats the relationship between markedness and the core/periphery distinction in the following way: Core grammar is said to be that part of the grammar generated by universal principles, for which the child needs only minimal exposure to acquire. In terms of markedness, core grammar represents the unmarked structures of the language. McLaughlin describes peripheral grammar as unconstrained by universal principles and as generated by natural changes in the course of a language's history as well as by borrowing from other languages. Peripheral structures thus are marked and must be learned.

Gass and Ard (1984) use the core/periphery distinction as an explanation for the likelihood of various linguistic phenomena to be transferred from L1 to L2. In particular, the authors state that language specific features, such as idioms, are less likely to be transferred, while core-like phenomena are more likely to show up in a second language. In addition to language specific features, Zobl (1984) includes aspects of a SL which are not "typologically consistent" with those of

the PL. According to Andersen (1984), Zobl restricts PL influence to those cases when the SL has unique, language specific, structural features or when SL features are not typical of the the genetic language group from which it derives.

These two positions are competitive with regard to the influence of PL in SL acquisition. Gass and Ard (1984) identify core grammar phenomena as candidates for transfer, while Zobl (1984) views PL influence as part of the periphery. Adjudication of this dispute is difficult at present given Kellerman's (1984) methodological criticisms of this line of research and Andersen's (1984) concurrence with them. In spite of the differences, however, questions about PL influence in SL use are no less relevant than they were in the heyday of Contrastive Analysis. Restricting the PL's role to peripheral aspects of the grammar may make pragmatic and semantic phenomena a more likely place to look for PL/SL interaction.

Universal Grammar, Parameter Setting, and SLA

Chomsky's (1981, 1986) theory of universal grammar (UG) incorporates a set of general 'principles' which apply to all grammars along with 'parameters' whose values are specified by the grammar of a particular language. Most of the work in the SL field conducted within this framework has focused on parameters and has included examination of branching direction (Flynn, 1984), pronoun-deletion (Phinney, 1987; White, 1984, 1985), pronominal reflexes, and subjacency.

In an overall assessment of UG, Cook (1985) concludes that "parameter fixing can formulate the relationship between first and second language learning in a more precise way" than previous attempts. Interference is said to result from access to UG by way of the first language, while SL production without interference is a function of direct access of UG. In a reformulation of the basic idea of contrastive analysis, instead of direct comparison of the surface forms of two languages, a UG approach to SLA examines differences in how two languages fix parameters for the same linguistic principle.

With regard to specific claims about the interaction of primary and second languages, Cook maintains that a PL influences those aspects of SL unaffected by maturation and those acquired rapidly, i.e. those aspects of SL which are similar to PL. This is more consistent with the position of Gass and Ard (1984) above, who see the major interaction between PL and SL focused in core grammar and not in peripheral features and structures.

A similar position is taken by Flynn (1984) in a study of Spanish and Japanese learners of English. Flynn examined the effects of principal branching direction on the processing of structures involving pre- and post-posed adverbials in sentences with and without pronominal anaphora. An elicited imitation procedure was employed. Results showed that subjects performed better when their first and second languages had similar branching directions (e.g., Spanish learners of English). When the parameters of a speaker's language do not match, however, as in the case of Japanese and English, Flynn (1984:86) concludes that the parametric values must be reset. In this study UG principles were found to operate across languages, while language specific parameter settings accounted for learner variation.

Phinney (1987), using the pro-drop parameter as a case in point, examined how the idea of markedness can be used to predict structural difficulties in learning a SL, at the same time addressing the more general question of why some languages appear to be easier to learn than others. Phinney states that L1 parameter settings represent unmarked, default options for the SL learner. These L1 settings are said to be generalized to L2, or even overgeneralized, until input from that SL brings about a need to reset the parameter. Delay in resetting will lead to production errors typical of SL learners. Data are presented from second language learners of both English and Spanish with regard to use of the pro-drop parameter.

Phinney's findings, as well as other research cited, suggest that when L1 is unmarked (e.g., Spanish in the case of pro-drop) and L2 is marked (English, which does not allow pronominal subject deletion), parameter resetting is difficult and time consuming. In the reverse situation (English learners of Spanish), going from a marked language to an unmarked one, parameter resetting was found to be easier, and even the beginners performed at a very high success rate. White (1984, 1985) reports similar data for Spanish and French learners of English, the former in effect transferring pro-drop properties over to English, the latter not.

Phinney's review of CA and EA is critical of these two approaches as being less of a theory than a way to analyze data. Her explanation and prediction of learner errors in the context of parameter resetting effectively reinstates the claims of the Strong Version of Contrastive Analysis.

All of these more recent approaches have a number of things in common. They are all theoretically motivated, either in linguistics (e.g., Greenberg's approach to linguistic universals or Chomsky's UG) or in language acquisition theory.

Furthermore, in the interlanguage tradition in which much of this body of research can be found, there is a common attempt to address cross-linguistic assymetries. Earlier attempts at CA compared surface structure features of the two languages; current approaches go beneath the surface, probing assymetries of principles and parameters.

Finally, whether due to the methodological tradition of CA and EA or out of principled motivation based on the particular research questions investigated, most of the work mentioned above reports on data from studies of production as opposed to comprehension, and many rely on uncontrolled or semi-controlled data collection techniques. We assume that in large part this is what led Kellerman (1984) to comment about variability in results and difficulty in drawing firm conclusions.

Preference Models of SL Use

In surveying much of the more recent work in SLA since the decline of CA and the rise of error analysis and interlanguage, Spolsky (1986) comments that "no one in the field would deny that a second language learner's knowledge of his or her first language has some effect on his or her performance in the new language" (p. 272). He presents the framework for what is described as a unified theory of SL learning based on Jackendoff's (1983) distinction between well-formedness conditions and preference rules. Spolsky's approach to SL learning is unique in that it incorporates social context conditions as the primary, or causal, elements of the model. Attitudes and motivation take their place alongside aptitude and previous knowledge (e.g., L1) forming what are described as "clusters of interrelated conditions or factors." Of greatest relevance here is that the various conditions are seen to interact in predictable ways; following Jakendoff, some take the form of necessary conditions, some are graded, and some represent typicality conditions.

Rutherford (ms.) cites evidence of typicality conditions in SL use from Ijaz (1986) on the perception of the meanings of locative prepositions and Kellerman (1978) on grammaticality preferences regarding various meanings of the word 'break.' Rutherford recasts a number of principles in SLA as candidates for preference rules, among them Andersen's (1983) "transfer to somewhere principle," Klein's (1986) set of principles to guide the early learner, and Kellerman's (1983) reasonable entity principle.

CHAPTER THREE: Integrative Second Language Processing

3.1 Towards Third Generation Models

Grosjean's Approach to Bilingual Research

Research in bilingualism, unlike the work reviewed above in SLA, generally followed the paradigm of Weinreich (1967) in looking for distinctions between single or compound bilinguals on the one hand and dual or coordinate bilinguals on the other. (This work is reviewed at length in Albert and Obler, 1978; McKormack, 1977; and Segalowitz, 1977).

A notable exception, or even reaction, can be found in the analyses of Grosjean (1982, 1985), who distinguishes two ways to address research on bilinguals. One he calls the "monolingual (or fractional) view;" the other he labels the "bilingual (or wholistic) view." In the former approach, bilinguals, their speech mode, and their manner of processing are divisible into two distinct "competencies." Research on fluency and proficiency, language skills and their social functions, and on the cognitive and developmental effects of bilingualism are interpreted by comparing bilingual subjects with monolingual counterparts. In contrast, the wholistic view speaks about an inseparability of the bilingual's two languages. Grosjean highlights this view by comparing the bilingual with the high hurdler. The latter, he states (1985:471), is qualitatively different from the sprinter and the high jumper just as the bilingual is not merely the combination of monolingual speakers of two languages.

Of greatest interest here is the way this wholistic view approaches the relationship between the two languages of the bilingual speaker. In Grosjean's words, there is a "co-existence and constant interaction of the two languages," which reflects a "an integrated whole, a unique and specific speaker-hearer." The primary area of research drawn upon for support of this distinction is the area of code-switching and borrowing. In this view of bilingualism, code-switching is seen as a unique language competence, rather than random or deviant as the fractional view of bilingualism would apparently have it.

It is the interaction of two languages during the course of second language learning and code-switching that are the foci of interest. In a more recent study Grosjean (ms) examines recognition processes involved in code-switching and borrowing, in particular the role of phonetic properties of host and guest words,

the point at which the switch apparently occurs, and language specific factors involved in the process. This study and a comment on code-switching among polyglot aphasics (Grosjean, 1985) illustrate an interest in the relationship and even interaction of the PL and SL in bilingual processing.

Second Language Reading Comprehension Research

Research in SL comprehension, like the work reviewed in the previous section, takes its lead from psycholinguistic studies in PL comprehension, in particular the field of reading (see especially the work of Carrell, 1983a, 1983b, 1983c, 1984 as well as the papers in Carrell, Devine, and Eskey, 1988). Three approaches were selected here as representative of a common thread of interest in SL users' processing of various levels of linguistic information (i.e., orthographic, syntactic, semantic, discourse).

Ulijn (1980, 1981) presents a model which distinguishes SL reading processes from those in L1. A primary concern is the role of L1 in the SL reading process. In particular, he names the limited nature of SL knowledge and PL interference in SL reading as constraints to be taken into account in the study of PL/SL interaction. He dismisses contrastive analysis, error analysis, interlanguage, and studies in bilingualism as irrelevant to SL reading.

In his model, Ulijn outlines an input component, a script recognizer, a sentence parser, a lexicon, and a conceptual system, focusing primarily on the role of conceptual and syntactic processes in SL reading. The conceptual system involves identifying content words and relating them to lexical and conceptual knowledge. In this category he also includes an inference mechanism which involves background knowledge and contextual information such as diagrams, numerical data, and other material outside the main body of the text. Syntactic processing also relies on information from the lexicon and the conceptual system.

Ulijn distinguishes his own approach from that of Clark and Clark (1977), who describe both meaning-driven and syntax-driven comprehension mechanisms. He seems to favor a conceptually-based system which utilizes syntactic information as a support or back-up for comprehension. Ulijn reviews a variety of studies, most of which favor a unidimensional approach to SL reading. For example, Aronson-Berman (1975) and Cowan (1976) see syntax as the primary source of difficulty in SL reading. Yorio (1971) regards lexical information as more relevant.

Ulijn's own research (Ulijn and Kempen, 1976; Ulijn, 1979) opts for a multidimensional approach, pointing out the role of PL-SL contrasts in vocabulary and syntax in SL reading comprehension. In particular, he reports greater difficulty in comprehension of contrasting lexical items in Dutch and French, but only a delay in comprehension with regard to syntactic contrasts. The appeal to syntax resulted only when the overall meaning was unclear.

Faerch and Kasper (1986) identify several hypotheses form L1 comprehension models which they explain as applying equally well to SL comprehension. Comprehension is defined as a matching process involving reliance on multiple sources of information, including communicative input, linguistic and world knowledge, and the linguistic context. Comprehension is said to include both top-down and bottom-up processes and inferencing procedures. It is also said to be both selective and partial. Two hypotheses, based on unique aspects of SL comprehension, are presented in this work. One relates to the SL user's limited knowledge in the second language, which leads to heavy reliance on bottom-up, text-based processing. A second hypothesis, ignored or even rejected in other models of SL use (see, e.g., Zobl, 1984) is that SL comprehension "often involves more than one language system" (1986:266).

Faerch and Kasper propose a model of SL comprehension which they say needs to encompass a concept of linguistic transfer, i.e. a way to distinguish interlingual and intralingual inferencing. The authors suggest a model based on competition between L1 and L2 and involving an innately specified program based on principles of universal grammar as well as modularity. Faerch and Kasper go beyond syntax, including lexical, pragmatic and discourse knowledge within their framework.

Sharwood Smith (1986), in a paper focusing on the role of comprehension as well as acquisition in a theory of language input, adopts a very strongly integrationist perspective. In a reinterpretation of Chomsky's competence/performance distinction which involves two kinds of knowledge as well as two kinds of performance (purely linguistic mechanisms to handle phonological, lexical, and syntactic relations and more general performance mechanisms to handle relations between the latter mechanisms and other competence systems). Sharwood Smith's model focuses in particular on interaction of existing (L1) performance mechanisms with target language competence (SL). In an example dealing with lexical knowledge, the model assumes that L2 competence (as limited as it is for the beginning learner) as well as L1 and extralinguistic information are all involved in making "inferences about L2 lexis" (1986:246). Sharwood Smith

distinguishes his interactive proposal from Krashen's (1981) rejection of L1 knowledge and Wode's (1981) ambivalence on the role of L1 in SL comprehension.

Autonomous versus Interactive Processing

In a controversy which spans the study of universals in first and second language acquisition, two positions have emerged. One school of thought favors an autonomous level of linguistic processing and can be shown to be based on Fodor's (1983) modularity approach. Clahsen (1984, 1987) provides the most detailed explication and strongest support for this perspective in the SL field. The other approach, dubbed a functional model, is exemplified by the Competition Model in the recent work of MacWhinney (1987a, 1987b), McDonald (1987), and Gass (1987) and is based on the earlier work of Bates and MacWhinney (1979).

Clahsen (1987) presents a model with two principle modules, including in his presentation counter-claims of functional researchers and a refutation of them. One of the modules is a grammatical processor (with subprocessors for handling lexical and syntactic information) which is said to map underlying representations onto surface strings. The other component is a general problem solver which can perform the same mapping function directly without invoking the grammatical processor. Two fundamental features of this model which Clahsen states distinguish it from functionalist approaches are the autonomous nature of the linguistic processor (what has come to be known as autonomous syntax) and the serial application of the different subcomponents of that module.

In a series of studies on the acquisition of German by adult foreign workers with Spanish, Portuguese, and Italian as first languages, Clahsen (1984, 1987) documents evidence for what he describes as "a strictly ordered developmental sequence" for the acquisition of major constituents in German word order. In the case of his Italian subjects, he explains that the first stage in the acquisition of German word order is identical to the underlying structure of L1 (Italian) word order. This fact is said to allow Italian learners of German to by-pass the syntactic processor, making use of the general problem solver for processing. Rather than transferring L1 surface structure word order to their second language, two of his Italian subjects chose to use canonical SVO patterns designated as the underlying representation for Italian, even though word order in German is highly flexible. Clahsen concludes that this transfer of underlying structure is evidence for an autonomous grammatical processor in SLA.

Kail (1987), in analyzing data from cross-linguistic studies of multiple cues in sentence and lexical processing, presents a somewhat different perspective, her model promoting notions of interaction as well as autonomy in linguistic processing. Picking up on a distinction between local and topological processing introduced by MacWhinney, Bates, and Kleigl (1984), Kail suggests that local processing may involve morphological and semantic information, while topological processing focuses on syntactic and pragmatic information (e.g., topicalization, focus).

The interaction idea is implicit in each of these separate components. It is claimed, however, that "simultaneous" processing of information at both the local and topological levels is impossible. This position favors the notion of autonomy, although somewhat different from the kind of autonomy in Clahsen's model.

In a series of studies describing their functional approach to language acquisition in general and to cross-linguistic studies in particular, Bates and MacWhinney (1981, 1982) present a performance model of sentence processing based on competition between different sources of information. In an application of the competition model to the study of bilingualism, MacWhinney (1987b) presents the ideas central to the model, discussing them within a context of transfer very much related to our interest here in integration processes.

The functional approach embodied in the competition model is introduced as distinct from competence-oriented, grammar-based approaches to acquisition and processing like those discussed above. The Competition Model (MacWhinney 1987a, 1987b) describes a process of direct mapping between form and function, where lexical, morphological, and syntactic cues are all integrated by the same processor. This differs from approaches distinguished by autonomous, modular processing components.

The relationship between form and function is treated by the model as interacting competitively during comprehension and production. A major concept in the model, cue strength, is the weighting of the connections between the various sources of information. This metric is said to permit statistical comparison between users of different languages with regard to the relative strength of cues such as animacy, number agreement, or word order.

Cue validity, another major construct of the model, is described by MacWhinney (1987b) as the most important for the study of SL learning. It is broken down into two parts: availability (defined as the ratio of the number of instances a cue

is available to the total number of cases) and reliability (the ratio of the instances a cue allows a correct decision to the number of cases the cue is available). These notions of availability and reliability are the basis for the notion of transfer in MacWhinney's (1987b) application of the Competition Model to bilingualism and SL use, and are said to account for approximately 95 percent of SL learning.

The remaining aspects of SL use, according to MacWhinney, require a dual system of cues. MacWhinney (1987b) suggests four patterns of cue strength, two of which involve integration of PL and SL information. These two are "transfer" of L1 cue weights to L2 and "merger" of those two systems. The non-integrative possibilities include "abandonment of L1 for L2" and "partial attainment of separate L1 and L2 systems" (1987b:322). Tendencies to transfer or merge cue strengths are discussed in terms of language history/relative age of acquisition of a bilingual's two languages, and the prevalence of code-switching in the environment.

In a study of both foreign and second language learners of English and Italian within this framework, Gass (1987) investigated the interpretation of sentences in which word order (syntax), animacy (semantics), and topicality (pragmatics) were manipulated. Subjects were asked to identify the subject of the sentence presented to them. In separate analyses of word order, animacy, and topicality, Gass interpreted the higher percentages of animate nouns selected (ranging between 71 and 82 percent, 18. <s.d.<24) and the greater homogeneity (standard deviation) in comparision to first noun selection (56 %-67%, 27<s.d.<32) as evidence for the greater importance of animacy in sentence interpretation.

Given this unifactorial design, Gass is forced to resort to a series of binary comparisons of word order, animacy, and topicality, resulting in the same kinds of problems raised below by the models employed by McLaughlin and Bialystok. Thus, the conclusion that subjects who go from a syntax-based language (e.g., English) to a more semantically-oriented one (e.g., Italian) do so with greater facility than those who move in the opposite direction may be premature, or may be in need of a methodology which can accomodate more than a single source of information.

3.2 Methodological Considerations

The models reviewed above were chosen because of their reference to information processing and/or their relevance to the study of SL comprehension. Two concerns which merit further elaboration emerge from this review, one theoreti-

cal and one methodological. The methodological issue refers to second and third generation models alone. Unlike the approach of the first generation models, which utilized production data as their source of evidence, the more recent approaches cited above utilize an experimental orientation as a means to test their proposals.

A general trend in second generation models seems to be a lack of consistency between the theory (both implicit and explicit) behind the model and the manner in which that theory has been operationalized. What appears to be a problem with regard to the distinctions between individual and experimental explanations as well as between unifactorial and multifactorial models in SL processing are discussed below.

The objective of the more recent proposals reviewed above is to predict SL performance expected as a result of learning. One of the apparent purposes of the Monitor Model is to predict the conditions whereby the SL user will invoke the monitor, thus indicating whether acquired or learned material has been used. In McLaughlin's information processing approach, 'learned material' is the main explanation for SL performance. Bialystok's model looks at SL use as a function of what the SL user knows and the degree to which that knowledge is accessible. Even though this orientation is not directly focused on acquisition, its appeal to information processing is somewhat less central.

Nevertheless, the descriptions of these models also express a concern for higher level processing in SL use like problem solving. This concern fits more properly into the domain of information processing than in the field of acquisition. The term processing appears frequently in the descriptions of these models, implying an interest in higher mental operations. Those higher mental operations cannot be implemented within the terminology of acquisition. This is because the study of acquisition attempts to predict how associations, habits, and the like are formed, maintained and retrieved.

An information processing approach attempts to investigate how already existing associations are processed. Thus, rules and types of processing fall within the purview of such an approach. Processing here is viewed as a phenomenon which occurs within the limits of time and space. The time constraints can be examined through an observational unit which sometimes continues for a number of minutes but more often is monitored only over the span of several seconds or milliseconds. An attempt to approach information processing issues in SL use makes it desirable to employ situational terms to meet descriptive and explana-

tory objectives. These terms need to be made more explicit in future models and paradigms to allow for validation of assumptions through experimental design.

Individual differences can help discover some characteristics of processing of interest. For others, however, and even for those aspects of processing for which an individual differences approach may be productive, situational factors ought to be examined as well in order to receive a more complete account of processing. As discussed elsewhere (Wolf and Walters, 1988), the comparisons made in Bialystok's (1984) model include both individual and circumstantial parameters. The subjects are defined as unilingual or bilingual. This aspect of the model is complemented by the use of two sorts of metalinguistic tasks (grammaticality judgements and sentence error correction). These can be seen as a manipulation of situational phenomenon. The combination of both individual and situational constructs in a single conceptual model is exemplary and rare (at least in the study of SL use).

Like the other models examined above, there is yet another requirement not fulfilled by this approach. This is the need for integrity between the overall methodological framework and the specific operational designs. More specifically, Bialystok's (1984a) research fits in the class of bifactorial models. In other words, the overall explanatory framework consists of two classes of relevant variables: analysis of knowledge and control of cognition. The assumptions of the design include those for main effects as well as an interaction. The former test for the independent effect of each one or both of the two variables; the interaction is a way to examine whether these two variables have some kind of combinatory effect. The bifactorial nature of the overall framework requires a design which gives equal entry to each factor. Such designs organize the combinations of the different levels of each variable (in the case at hand, 2 X 2 = 4 conditions) in a single, unified framework, affording each one the same prerequisites with regard to instructions, task, measurement scale, and the like.

The most desired model to carry out this set of experimental requirements is a bifactorial design. Bialystok's (1984) model takes advantage of the conceptual aspects of such a design, in its incorporation of both individual and situational variables. Methodologically, however, we do not find a single bifactorial design with four, equal conditions. The task variables (grammatical judgements and sentence corrections) as well as those of text (anomolous and nonanomolous sentences) are nested within the basic bifactorial design. The comparisons, then, are made between non-equalized conditions. From a strictly methodological

point of view, this means that the actual explanatory attempt was unifactorial and not bifactorial as intended.

3.3 Substantive Issues

A major theme which emerges from first, second, and third generation thought in the field of SL studies has been variously described as borrowing and interference, positive and negative transfer, mixing and code-switching, and interaction and integration. The common element in this work is the role of PL information and knowledge in second language use. In the early studies the focus is more on production, or speech, in a second language; in the later work interest centers more on processing. In the early studies linguistic analysis is primary; in the latter ones psycholinguistic issues are more central.

Beyond generational differences, all of this work shows a major concern with the role of the first language in SL use. In borrowing and interference as well as in transfer, the relationship between the PL and SL is assymetrical and often unidirectional. In code-switching and integration the PL may have a more balanced role with the target language. If we combine these positions with a null hypothesis, i.e. that the first language has no role in SL use, we are left with three fundamentally different proposals. The PL has no role in second language use; the PL has a deleterious influence on SL use; and the PL has a contributory effect on SL use. The experiments presented in Part II are an attempt to address these hypotheses.

3.4 More on Integrative Processing

It is well-accepted (but see Krashen, 1982, and Zobl, 1984) that any new SL linguistic system operates on the basis of generalizations from an already acquired and functioning PL system. SL processing is assumed to rely to some extent on PL knowledge. More specifically, the SL user searches for similar PL linguistic units (nodes, properties, etc.) and integrates information from both languages in order to comprehend in his second language.

The principle idea of the CAH, a first generation theory which deals with SL use, can be reexamined and reconstructed here. Inasmuch as SL linguistic material matches its PL counterparts, SL performance is expected to be effective. A mentalistic implication can be derived here, i.e. that SL use can be characterized by a process of comparison between the features of SL material that the user is

exposed to and the relevant aspects of his already established PL. This restatement of the CA proposal provides a theoretical basis for our assumption that PL knowledge is incorporated into SL processing.

Following this logic, the 'contrastive' notion of that first generation theory is revived by means of a central construct implied in second and third generation models, i.e. integrative processing. The term 'contrastive' implies a process of comparison between two linguistic systems in order to validate one of them, namely, the SL. In that process, new information from SL is afforded meaning through an integrative procedure. It can be viewed as if the SL elements which are compatible with PL components are integrated with the latter by the process of SL comprehension. The integration operation may add compatible elements; it may eliminate incompatible SL elements from those in the PL; or, some midpoint (averaging) between the old and novel pieces of information may be established.

From a methodological perspective, the literature review treated the problem of fair and balanced empirical comparisons. This problem tends to emerge when a conceptual framework is not accompanied by an appropriate method. To allow tests for the independent role of PL and SL as well as a test for integrative processing (interaction, in classical terms), a full factorial design is considered appropriate.

'Latent' factorial models (e.g., those which do not apply a complete multifactorial design for an empirical examination of a multidimensional conceptualization) imply an untested assumption that the SL comprehension process involves more than one explanatory element. These models cannot test for the interactive model embedded within the overall conceptual framework, since they do not programmatically and systematically use complete factorial designs. Consequently, they are unable to treat questions addressing the integrative nature of SL comprehension.

Assuming an adjustment of a complete factorial design to the requirements of factorial modeling, there is still a need to resolve another fundamental question. A factorial design conventionally tests the assumption that the relevant factors interact with each other. However, the drive to clarify the nature of SL processing remains unfulfilled, since information regarding the statistical significance of the interaction term does not convey theoretical meaning with regard to integrative aspects of that process. To illustrate this point, assume that using a factorial design for the study of integrative aspects of processing, a researcher finds two main effects and no significant interaction. Would he conclude that the two

perceived pieces of information are not integrated within the reader's mind? Conventional factorial modeling does not allow a rejectable test of such a conclusion (for more details, see Wolf and Walters, 1988).

With regard to prevailing 'interactive' approaches (Anderson, 1981; Kintsch, 1988; Kintsch and van Dijk, 1978; Laberge and Samuels, 1974; Lesgold and Perfetti, 1981; Rumelhart, 1977; Rumelhart, McClelland and the PDP Research Group, 1986), a potentially clear distinction can be drawn between an attempt to account for the mechanism of language use and an attempt to induce the rules that govern processing. The former case provides a mechanism for a dynamic and detailed description of the way information is processed, like design specifications on an electrical circuit board. It is not a coincidence that such models go hand in hand with neural research into processing. In a rules-induction approach to processing, the interest is in formal specification of the nature of the process.

Our choice between these two alternatives was motivated by an apparent need in the field of SL comprehension to identify focal cognitive units which may play a role in that process. Information Integration Theory (Anderson, 1981) seems to fill this need. Information Integration Theory (IIT) is based on the assumption that "...all thought and behavior has multiple causes, being integrated resultants of multiple sources of stimulus information" (Anderson, 1981:4).

IIT has been applied to a number of problems in psycholinguistics (e.g., Anderson, 1981; Greuneich, 1982; Leon, 1982; Oden and Anderson 1974; Oden 1977, 1978; Walters and Wolf, 1988; Wolf, Walters, and Holzman, 1989). Much of this psycholinguistic research with IIT has attempted to examine the relationship between various parts of sentences. One study in particular has shown how different pieces of semantic information can be integrated into a single judgment (see Oden and Anderson, 1974).

The paradigm presents the reader with sentences, each of which contains a combination of a particular level of one source of information with a particular level of a second source. The reader responds to these sentences consecutively on a rating scale. The complete design is factorial, i.e. it includes all combinations of the levels of each source of information. The data are submitted to both descriptive and inferential statistical analyses in an attempt to clarify the importance assigned during the SL comprehension process to the different pieces of information and the rule of integration which relates these weighted percepts to each other.

The stimulus information experimented with has generally included different bi-factorial combinations of a person's sociableness, interestingness, trustfulness, dependableness, gregariousness and pleasantness (Oden and Anderson, 1974). It has also included work on the integration of intentions with consequences and regrets (Grueneich, 1982; Leon, 1982); relative truthfulness of premises involving category membership (Oden, 1977); and relative sensibleness of the readings of ambiguous sentences (Oden, 1978).

The results from studies conducted within the IIT paradigm points to the multifaceted nature of the processes under investigation. Regarding the type of integration rule, multiplicative rules have been documented in experiments on attribution and person perception, while additivity was the finding in studies involving sensibleness evaluation.

Functional Measurement (FM), the methodological counterpart of IIT, is introduced by Anderson (1981:12) as a reversal of traditional practice, making "measurement theory an organic component of substantive investigation." According to the guidelines of functional measurement, the measures and sampling procedures are valid only if the observed functions follow a predictable trend which can be formulated in algebraic terms. An example of FM is presented here for illustrative purposes.

In Oden and Anderson (1974) subjects were shown sentences which factorially manipulated agent-verb and verb-recipient constraints. One of the target sentences read as follows: 'Mr. X is very sociable; how likely is it that he will socialize with very sociable people?' The sentence here consists of two sources of information, one about the agent and one regarding the recipient. Four levels of each source of information yielded a total of 16 trial sentences. For each trial a different combination of agent and recipient information was presented, the subject making judgments of likelihood on a 20-point rating scale. The example above is expected to elicit a high likelihood of sociableness. The following sentence, which contains a weak complex of information, ought to elicit a low likelihood: 'Mr. X is very unsociable; how likely is it that he will socialize with very unsociable people?' A third trial from among the 16 included a high level of agent sociableness along with a low level of recipient sociableness, as follows: 'Mr. X is very sociable; how likely is it that he will socialize with very unsociable people?' In a trial of this sort, if the subject takes into account the sociableness of both agent and recipient information (as he did in Oden and Anderson's study), he gives a likelihood rating somewhere in the middle of the rating scale.

The entire set of trials was composed of all the factorial combinations of agent-verb and verb-recipient information, as follows: 'Mr. X is very sociable (or moderately sociable, moderately unsociable, very unsociable); how likely is it that he will socialize with very sociable (or moderately sociable, moderately unsociable, very unsociable) people? The agent-verb constraint in this sentence is represented by the noun-adjective combination (Mr. X is very sociable, etc.) and the verb (socialize), while the verb-recipient constraint is based on a combination of the same verb and the adjective-noun complex describing a second character (very sociable people, etc.).

The graph in Figure 3.1, taken from Oden and Anderson's (1974) findings, illustrates the different kinds questions FM is capable of handling.

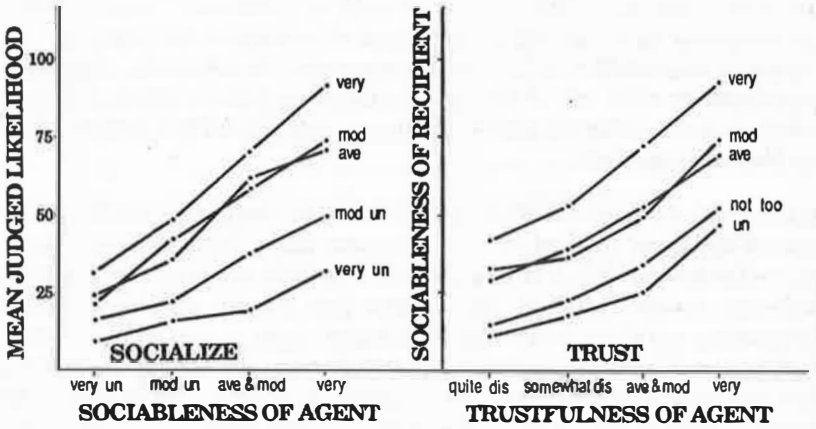


Figure 3.1. Mean judgements of likelihood that agent will socialize with recipient (from Oden and Anderson, 1974:140).

A survey of the slope of the curves in Figure 3.1 from the lower-left to the upper-right-hand corner of the graphs reflects the impact of the 'recipient' information, while the distances between the various curves indicates the impact of the 'agent' information. The fan-shaped pattern in the left panel represents a differential weighting valuation strategy. A left to right survey of the curves

indicates that, for any level of agent sociableness, the distance of mean judged likelihood related to the corresponding level of recipient sociableness is greater. Thus, the increase in judged likelihood as a function of both agent and recipient information is greater than the simple summation of them.

FM is proposed as a possible technique for examining questions about comprehension processes in a second language and other situations of partial linguistic uncertainty such as understanding of metaphors and idioms. Application of FM requires construction of SL sentences or texts such that both PL and SL information can be manipulated. In this manner, the general question to be addressed is whether PL information is integrated with SL information during the SL comprehension process?

PART II: EMPIRICAL EVIDENCE

CHAPTER FOUR: An Introduction to the Method and a First Pass

Traditionally there are three levels for carrying out the analysis of linguistic and psycholinguistic issues — the text level, the sentence level, and the word level. Teachers of English as foreign language prefer the first, or text level, for teaching and testing purposes. This preference is expressed in the form of an 'unseen' passage with questions on each of the three levels, according to the individual teacher's needs. Such an approach is indeed reasonable, since academic ideas, no matter what level of abstraction, need more than a single sentence to be introduced and developed. It is also more economical to display a testing framework capable of including all three levels of analysis at the same time than to examine them separately.

As a first, exploratory step towards the establishment of an appropriate mode of response for our empirical work, we adopted the above-mentioned solution and selected short narratives as the starting point for measurement. In approaching the comprehension task, again in exploratory fashion, we chose to try out two methods. One was a more or less direct question about information in the narrative. The other was a metalinguistic judgement based on knowledge of texts in general and the experimental text in particular.

Unfortunately, it may not be possible to manipulate bilingual information directly at the text level of analysis. We had to take proficiency differences as a point of departure for the selection of a mode of response appropriate for collecting valid data on integrated second language use. This choice is based on the simple fact that language proficiency is the most salient aspect of SL use. English as a foreign language teachers throughout the world naturally classify both students and teaching materials as advanced, intermediate, and beginner. Viewing proficiency as a key to substantive psycholinguistic differences, this approach involves the underlying assumption that processing routines, capabilities, and learning schedules are unique to a given proficiency level.

IIT approaches psycholinguistic processing in terms of subjective valuation of cognitive units of information embedded in a linguistic context. The next step of processing is application of an algebraic rule for the integration of those percepts.

Accordingly, we expect valuation and/or integration differences between language users of different SL proficiency levels.

There are two contradictory intuitive working hypotheses. On the one hand, proficient SL users may process a full range of (in our terms, both semantic and structural) information when they perform a metalinguistic task. That is because it is easier for a proficient language user to go beyond the dictionary meaning and beyond grammatical shortcomings in ability. On the other hand, proficient subjects might lose their comparative advantage precisely because a metalinguistic task overshadows their relatively superior experience. Either outcome is satisfactory, since it must point to either a metalinguistic or direct comprehension task as a source of proficiency-related differences. The alternative finding of no valuation differences should indicate a failure in our experimental gamble.

A series of experiments, one of which was reported elsewhere, addresses the exploratory issue raised here. In these experiments IIT was used to examine whether semantic and structural information might be integrated by some regular algebraic rule and whether or not the experimental task might influence the integration process.

Experiment 1
Integration of Text Information in Fluent SL Readers

In an experiment reported on elsewhere (Wolf, Walters, and Holzman, 1989), twelve female high school students who were fully fluent Hebrew-English bilinguals were tested. All were members of a special tenth grade English class for native speakers of English. The design involved manipulation of semantic and structural aspects of the same narrative passage. Semantics was manipulated in the texts by varying the logical compatibility of an industrious boy and his report of having completed all, some, or none of a work assignment. Three levels of compatibility were examined. Structure was manipulated by varying the order of elements in a story grammar framework (Stein and Glenn, 1979). Three orders resulted from presenting the narratives in a canonical order, a partially-mixed version, and a fully scrambled fashion.

The canonical order of the narrative used was manipulated by introducing a very industrious volunteer on a kibbutz (a communal agricultural settlement in Israel). Within the narrative, the volunteer was asked to water flowers in all of the kibbutz gardens. In his attempt to fulfill his supervisor's request, the volunteer goes out to the gardens with a watering can. Next, he reports to the supervisor on his

efforts. The partially-mixed version involved presentation of the above story grammar categories in reverse order, while the fully scrambled narrative was an arbitrary order of all categories except for the setting sentences, which always appeared first. Nine narratives resulted from a factorial combination of the three levels of semantic compatibility with the three levels of structure. Subjects were tested individually, using a 20-point rating scale.

To examine the role of task, two kinds of comprehension questions were asked in different experiments. A metalinguistic question asked subjects to make a judgement about the logic and organization of the narratives. Logic was intended to refer to meaning-related elements of the text, while organization was explained to refer to structure-related properties. A direct comprehension question related to text-based information was the task manipulation in the second experiment. Subjects were asked to rate the likelihood that the protagonist really did the work assigned. The graphs in Figure 4.1 present the data pooled across the 12 subjects for the metalinguistic task (left panel) and the direct comprehension question (right panel).

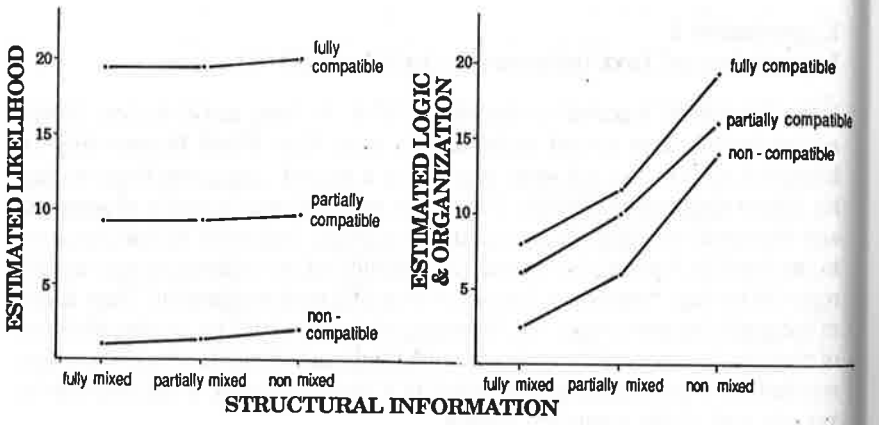


Figure 4.1 Mean estimate of likelihood (left panel) and mean estimate of logic and organization (right panel) for 12 bilingual subjects (after Wolf, Walters and Holzman, 1989).

In the graph above, the semantic information is represented as the different curves, structural information is marked along the horizontal axis, and the ratings of the subjects are plotted on the vertical axis. Distance between the curves is taken as indicative of an effect for semantic information, i.e. that subjects relied on semantic aspects of the text in making their estimates. In corresponding fashion, the slope of the curves is a measure of the contribution of structural information to the subjects' ratings. Finally, the pattern of the curves (e.g., parallel, fan-shaped) represents valuation and integration typicalities.

In the direct comprehension task (left panel of Figure 4.1), semantic information was relied upon almost to the exclusion of structure. The wide spacing between the curves attests to this finding. The lack of slope in the curves and the almost identical marginal means (10.4, 10.0, and 9.9, respectively) for structure indicate basically unidimensional semantic processing for the direct comprehension task. A two-way ANOVA for repeated measures yielded a significant main effect for semantics; structure was found insignificant.

When bilingual subjects were asked to undertake a metalinguistic task, rating the logic and organization of the narrative, a somewhat different picture emerged. As can be seen in the right panel of Figure 4.1, there is well-spaced distance between the curves as well as a noticeable lower-left to upper-right slope. Main effects for semantics and structure were significant, indicating that both were taken into account in subjects' judgements in the metalinguistic task.

The parallel pattern of the curves indicates that the two sources of information were integrated by a simple weighting rule. (The insignificance of the interaction term corroborates this). It seems, then, that under conditions of metalinguistic comprehension, fully fluent subjects estimate the value of both semantic and structural information in a text while in direct comprehension only semantic information is relied upon.

Experiment 2
Integration of Text Information in Non-native Readers

In an attempt to replicate the above work on less proficient second language users, 18 Israeli high school students, speakers of English as a foreign language, were sampled. This group was equated with the bilinguals mentioned above for relevant sociological and psychological parameters. The students were exposed to the same stimulus narratives and were asked to respond in the same two tasks,

one direct comprehension and one metalinguistic. Figure 4.2 presents the data pooled across all 18 subjects.

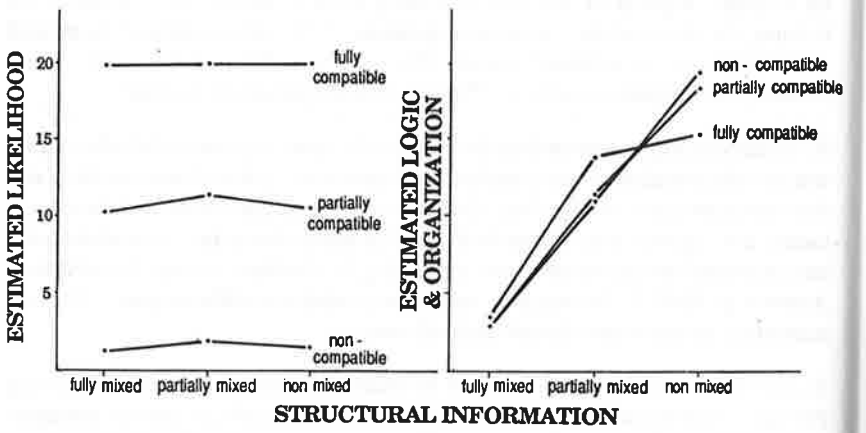


Figure 4.2. Mean estimate of likelihood (left panel) and mean estimate of logic and organization (right panel) for 18 non-native readers of English.

There is no difference between the shape of the graphical structure in the left panel in Figure 4.2, which represents direct comprehension, and the respective pattern in Figure 4.1. Both clearly indicate the strongest possible effect for semantic compatibility alone. The large F-ratio for semantic compatibility here, $F(2,17)=1503.01$, along with the non-significant F-ratio (2.49) for structure strongly support the above impression. This graphic similarity implies that both the valuation and integration aspects of the comprehension process were alike for proficient and relatively non-proficient SL users. This finding permits a comparison of these processes under the same conditions in the other, rather different, metalinguistic task.

In the right panel of Figure 4.2 the marked slope of all three curves and the proximity among them indicate that only structure was taken into account for the processing of a metalinguistic task. The structure effect was significant, $F(2,15)=156.34$, while the semantic effect was not, $F=1.63$. The narrowing of

cognitive scope seen here (reliance on only one source of information) is not consistent with the wider scope displayed by proficient SL users under the same stimulus and task conditions, as seen above in the right panel of Figure 4.1. There, proficient users related to both substantive elements in the text, i.e. semantics and structure.

It is puzzling why the non-proficient subjects assigned so much importance to the structural information in the narrative, practically ignoring the semantic aspect, which they had preferred in the direct comprehension task. Was it the very nature of the task that completely shifted their valuative attention from one linguistic domain to another? The question is open to further conceptual and methodological analysis. As stated, however, the entire picture points to the utility of a metalinguistic comprehension task for the study of integrative processes in second language use. Thus, in further experimentation we adopt metalinguistic tasks for an examination of an integrative approach to SL comprehension.

As mentioned above, the text level of analysis does not lend itself to simultaneous manipulation of both PL and SL information. To accomodate both sources, the two more constrained structural levels of analysis, the sentence and the word, seem more appropriate. The following chapters describe several series of experiments along this recommended line.

CHAPTER FIVE: Metaphor Comprehension

Metaphor, as a linguistic product, is special by virtue of its attempt to convey uncertainty rather than striving for precise representation of literal meaning. In most formal, linguistic respects, metaphor has nothing in common with the field of second language use, especially since metaphors, like idioms, are language-specific and not often translatable. For this reason, however, metaphors (and idioms) present a unique challenge to a second language user. That user is not generally familiar with informal, pragmatic forms of expression. Thus, studies of metaphor may provide entry into the nature of coping with uncertain linguistic information typical of second language use. We focus here on metaphors in two languages, Hebrew and English, which function as first and second languages in the context in which the studies were conducted. We sampled Hebrew and English native speakers, providing them with pure metaphors in these languages as well as metaphors containing bilingual information.

The use of metaphor is widespread, appearing in ordinary speech, dialogues, narrative prose, poetry, and sometimes even in academic writing (Lakoff and Johnson, 1980). This subject was recognized even in the early days of civilization by Aristotle, who noted in 'Poetics' that the use of metaphor is a sign of inborn talent and that good use of metaphor is identical to theoretical thought. The very term he used to label his subject (Greek, 'tometaphorikon' = the ability to use metaphor) is closer to a psychological perspective than to a pure linguistic approach. These two perspectives are of relevance to the studies which follow, since the former is the one we prefer to examine, while the latter is the most investigated one.

The linguistic/philosophical treatment of metaphor includes: the distinction between literal and metaphoric forms (Fraser, 1979; Ortony, 1979, 1980; Rumelhart, 1979); the role of similarity and comparison (Black, 1962, 1979; Perrine, 1971; Searle, 1979); the relationship between the basic parts of a metaphor, i.e. topic/ vehicle (Richards, 1936); the distinction between simile and metaphor (Fraser, 1979; Kintsch, 1974; Miller, 1979; Ortony, 1979); and the role of speaker intentions (Fraser, 1979; Searle, 1979).

The same issues have been addressed from a psychological perspective. In addition, psychologists have examined: how a novel entity emerges from disparate parts (Billow, 1977; Paivio 1979; Winner, 1979); the role of linguistic and extralinguistic context in comprehension (C.C. Anderson, 1964; Honeck et al.,

1980; Paivio, 1979; Ortony, 1980); memory of metaphors (Paivio, 1979; Harris et al., 1980); analogic reasoning in metaphoric understanding (Fraser, 1979; Miller 1979); and the role of imagery in metaphoric processing (Paivio, 1979; Reichman and Coste, 1980). The latter subjects are closest to our locus of intention in their consideration of processing. However, it is not what we mean when we speak of the nature of processing.

Fraser (1979) defines metaphor "...as a predicating expression which is used non-literally, which requires the establishment of an analogy for its successful interpretation." Considering this definition, metaphor understanding represents a specific case of going beyond the information given. When a person hears a literal statement, he can understand it by decoding the statement into its syntactic and semantic components. For example, the statement "Where is your book?" uttered by a woman to her husband would not provoke controversies and meaning discrepancies, since its meaning can be directly derived from the statement itself (which is a query about the location of the listener's book).

When a person uses a metaphor, however, he is using a non-literal statement which requires the listener to interpret the meaning of the sentence based on external information. For example, the sentence "John is a wolf" cannot be understood literally since it violates the categorization of animals into human and non-human beings. Instead, this statement denotes another meaning that the hearer could understand by looking for external information, such as known properties of John and wolves, situational context, the speaker's intentions, the relationship between the speaker and John, etc. In fact, the hearer can give different meanings to this statement, e.g. "John is a cruel person" or "John is a fearless person." Thus, the challenge to the hearer is to figure out which of the possible meanings the speaker has in mind. This choice of meanings is permitted by the integration the hearer performs on collected information regarding the uttered expression, its components, and the speaker's intentions. Unfortunately, there is no solid evidence regarding the processes by which a metaphor is understood in this way.

Psycholinguistic literature has been more concerned with the definition and philosophical implications of metaphor as a special linguistic phenomenon. One of the theoretical frameworks for investigating metaphor was developed by Fraser (1979) based on a model presented by Miller (1979) which maintains that metaphors derive from similes by systematic deletion and rearrangement of certain parts of the statement. According to Fraser, the interpretation of a metaphor can be described as the decoding of the metaphorical expression

(vehicle) into its canonical simile and the comprehension of the properties that were compared. This allows an analogy to be drawn between the literal and metaphorical elements. Put simply, if we hear the sentence "John is a wolf," our main task is to discover which properties of John and wolves were compared by the speaker. If, for example, the speaker has in mind that John is cruel like a wolf, the speaker has equated John with a wolf (through the shared property of cruelty). In this case, the most correct interpretation of this sentence would be "John is a cruel person." However, the hearer cannot discover this meaning by simply hearing the sentence, since he cannot be attuned to the speaker's intentions and analogies.

In order to choose a meaning, the hearer must collect information about the possible properties of the different components of the simile or metaphor. For the work described below, insight into cognitive processing in general and metaphoric understanding in particular does not lie in what is processed but in how the human mind operates when challenged by the tension between literal meaning and intended meaning. This is the metaphoric dilemma as we see it. Questions like the following emerge: Is the process a simplified one, i.e. involving only one source of information? Or, is metaphoric understanding a complex process, involving more than one source of information integrated into an observable verbal response? Two sources of pragmatic information are involved in metaphors. Topic conveys the subject of the metaphor; vehicle provides the listener a context through which to channel the interpretation in a non-literal direction. A working hypothesis states that both sources of information modify the way the listener accomodates the meaning of a metaphor.

Experiment 1
Comprehension of a Hebrew Metaphor Involving an Abstract Topic by Hebrew Native Speakers

Methodological Specifications: Subjects and Materials The present experiment was conducted to examine integration of topic and vehicle information in the comprehension of a relatively obscure Hebrew metaphor.

Ten Hebrew native speakers, ranging in age from 35 to 50, all of whom were college graduates, half male and half female, participated in this experiment. The following metaphor, which expresses exaggeration and is literally impossible, was selected: 'X is nothing but a flying camel.' This metaphor was chosen because it is not widely used in daily conversation. Thus, it conveys a degree of

uncertainty for the language user. The place of the topic of the metaphor (X in the above statement) was filled with the following content: 'The desire for economic independence...' This generated the complete metaphoric sentence "The desire for imminent economic independence is nothing but a flying camel." In Hebrew: 'ha-she'ifa l'atsma'ut kalkalit b'karov eino ela gamal poreach b'avir.' (We note here that this or any English translation cannot capture the full conveyed meaning of the given metaphor.)

Four abstract nouns expressing different degrees of desire were selected. These lexical items were ranked in terms of strength of desire, as follows: 'kmiha' (longing), 'mishala' (wish), 'she'ifa' (desire), and 'tikva' (hope). These nouns were intended to set up four levels of topic information to be combined with four levels of vehicle information. Since we are dealing with exaggeration, we selected three animals (in addition to 'camel') of distinguishably different size to form the following set of vehicle substitutes: elephant, camel, rabbit, and mouse. The four levels of topic and four levels of vehicle information were combined in a complete factorial design, yielding sixteen different metaphoric sentences. Each sentence was followed by a question about how far the state is from economic independence.

Sixteen cards were prepared, each containing the following information: One of the 16 metaphoric sentences, the comprehension question, and a graphic rating scale. Two complete 3 X 3 factorial combinations of distinct metaphoric sentences were used for warming-up purposes. The practice sentences were as follows: 1."The (model/actress/princess) has (silk/velvet/cotton) skin. How soft is her skin?" 2.Encyclopedias/journals/dictionaries are gold/silver/ copper mines. To what extent does it serve as a reliable source of information?" The subject made his ratings on a 20-point scale.

Procedure. Each subject was tested individually. The entire procedure took approximately 45 minutes. Following the warming-up period, each subject was presented with the sixteen stimulus sentences in an arbitrary sequential order, rating how far the state was from economic independence.

All individual graphic patterns showed similar trends, and therefore are pooled for presentation here, as can be seen in Figure 5.1.

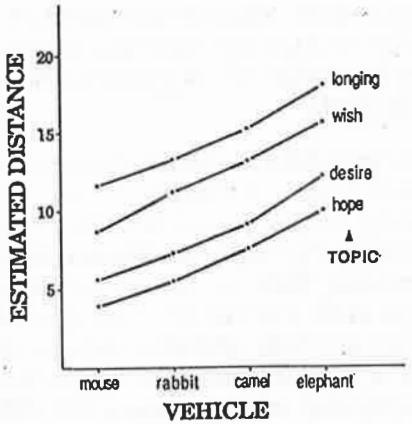


Figure 5.1. Mean ratings of distance from economic independence as a function of vehicle and topic information in 10 native speakers of Hebrew.

Both vehicle and topic were relied upon by subjects in the comprehension task, as can be seen by the slope of the curves and the well-spaced distance between them, respectively. An ANOVA indicates a noticeably stronger effect for vehicle as compared to topic, F(3,27)=89.91 versus 10.35 (p). One should, however, refrain from interpreting this difference in weighting without relating to the observed vehicle advantage as a reflection of the fact that the only way that a topic can gain meaning is through the vehicle.

A visual inspection of the relative slope of the four curves shows parallelism, indicative of a simple weighting valuation strategy. This is supported by the insignificance of the interaction term, F(9,81) Anderson and Lopes (1974) manipulated adjective by adjective by noun stimulus combinations in order to test for the precise valuation strategy and integration rule in the judgement of job proficiency. To demonstrate the use of an averaging rule of integration between the two adjectives, the authors expected (and found) graphic parallelism between these two elements. To show that adjective valuation depends on the noun, they expected (and found) interaction between these two sources of information. Due to the functional unity of vehicle and topic information in a metaphor, it would

be inappropriate to generate an empirical test to clarify the precise nature of the rule of integration, as might be advisable for exploratory work using functional measurement.

Anderson and Lopes (1974) manipulated syntactic elements that do not belong altogether to any legitimate phenomenological unity such as metaphors, idioms, etc. Thus, they were able to generate stimulus combinations as required for precise determination of the integration rule. While we seem to suffer from an inability to attach elements external to topic and vehicle in order to carry out a test for the integration rule, we do not have good reasons to deny the inherent similarity between the psycholinguistic connotation of Anderson and Lopes' work and our exploratory experimentation on metalinguistic comprehension of metaphor. As a working assumption for the interpretation of our findings, we adopt Anderson and Lopes' conclusion with regard to the parallelism found in their study as pointing to the application of an averaging rule of integration.

With regard to the parallelism of the results, indicative of the use of simple weighting, it is somewhat early to decide whether the application of this strategy is constrained by stimulus information (i.e., syntactic and semantic properties) or by personal information (language background). The following experiments attempt to establish a data base to help clarify this issue.

Experiment 2
Comprehension of a Hebrew Simile Involving an Abstract Topic by Hebrew Native Speakers

The above preliminary conclusions are limited to a context in which Hebrew native speakers comprehend a Hebrew metaphor in which an abstract topic is manipulated along with vehicle information. Experiment 2 was conducted to try to establish some generality for the findings of Experiment 1. A variation in the content of the metaphoric expression was selected as follows: "The love between man and woman is as strong as death." The marker 'as' prior to the vehicle information is traditionally what distinguishes simile from metaphor. Following Fraser (1979) and Miller (1979), we consider metaphor and simile to be pragmatically part of the same phenomenon.

Methodological Specifications: Subjects, Materials and Procedure

Ten sixth grade students, native speakers of Hebrew, all high achievers, half male and half female, participated in this experiment. One subject was unable to adjust

to the task and was excluded from the sample. The target sentence, including four levels of topic and four levels of vehicle information, was as follows: "The love between (man and woman/father and son/brother and sister/boyfriend and girlfriend) is as strong as (death/illness/a wound/a scratch)." The four former alternative social pairings represent different levels of an abstract topic, which is the strength of the emotional bond, while the latter four vehicle terms portray analogs of the strength of the topic. These terms were combined in a complete, 4 X 4 = 16, factorial design. Subjects were asked to rate the strength of the love expressed in the simile. The procedures, including the warm-up routines, the rating scale and all other related aspects, were identical to those in Experiment 1.

Results and Discussion

Due to the homogeneity of the individual ratings, the data from all 9 subjects were pooled and are presented in Figure 5.2.

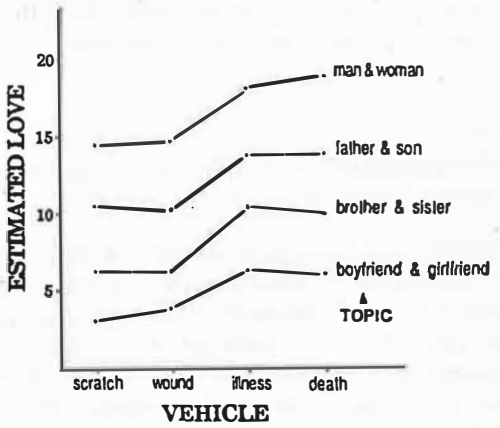


Figure 5.2. Mean ratings of strength of love as a function of vehicle and topic information in 9 native speakers of Hebrew.

The overall picture here is similar to the one observed in the previous experiment (see Figure 5.1). Vehicle and topic information were both relied upon, as can be seen by the slope of the curves and the distance between them, respectively, both

in the predicted direction. As was the case above, vehicle information captured most of the subjects' valuative attention, $F(3,8)=90.49$, as compared with topic, $F=7.12$ (p).

We again refer the reader to our cautious comment with regard to relative weighting in the previous experiment. Rather than providing evidence of subjective preference for vehicle information, the replication of the valuative advantage of that vehicle information in this experiment should be related to as a reflection of the total functional dependence of the topic's meaning on the conveyed meaning of the vehicle. Careful examination of the marginal means reveals that topic information was collapsed into two categories, two lower levels (boyfriend/girlfriend and brother/sister, 8.65 and 8.79, respectively) and two upper levels (father/son and man/wife, 12.26 and 12.32, respectively). Again, the valuation strategy was one of simple weighting, as implied by the parallelism of the curves and the insignificance of the interaction term, $F(9,72)=1.06$.

By now there is an observed valuative invariance beyond syntactic and semantic constraints such as the grammatical category and abstractness of the topic and vehicle terms. On this basis a preliminary impression can be formed, namely, linguistic differences may not account for integration processes in the meta-linguistic comprehension of metaphor.

Experiment 3
Comprehension of a Novel Metaphor Involving a Concrete Topic by Hebrew Native Speakers

The metaphors in the previous experiments are rooted in the language and culture of the subjects. Here, a novel metaphor was generated for experimental purposes. It consists of a concrete noun topic and a vehicle with a noun phrase formed by a derived adjective and a head noun, as follows: "The politician entered the room, and people whispered, 'Here comes the sewer of words.'"

Methodological Specifications: Subjects, Materials, and Procedure

Eight college-educated, adult women, ages 32-34, all married with children, and native speakers of Hebrew, participated. The stimulus material consisted of a metaphoric sentence with four alternative topic terms and four alternative vehicle expressions. The sentence with all stimulus alternatives was as follows: The (politican/ journalist/teacher/writer) entered the room, and people whispered,

'Here comes the (scribbler/sewer/tier/rhymer) of words.'" All possible factorial combinations yield a total of 16 sentences.

The subject was presented with each sentence and asked to rate the sensibility of each one. The entire procedure was similar to that reported in Experiments 1 and 2.

Results and Discussion

Figure 5.3 graphically depicts the data for the eight subjects in Experiment 3.

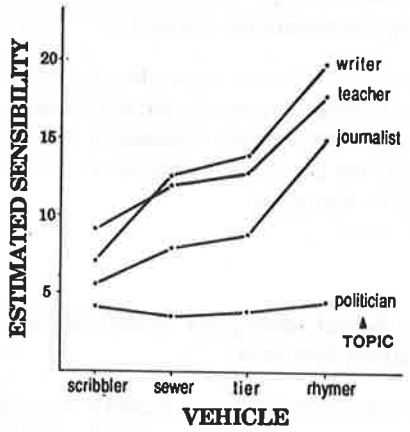


Figure 5.3. Mean ratings of sensibility as a function of vehicle and topic information in 8 native speakers of Hebrew.

As in the two previous experiments, both topic and vehicle information were relied upon, with a decided advantage for the latter, $F(3,7)=7.80$ and 32.10, respectively (p). The valuation strategy seems, however, to differ in the present study, as implied by the fan-shaped graphical pattern. This can be seen in the linear divergence of the curves from the lower left to the upper right hand portion of Figure 5.3. This graphic impression is supported by the significant interaction term, $F(9,63)=10.80$, p.

If the Topic X Vehicle model fits the trend of the data, then the interaction should be concentrated primarily in the bilinear component, and the residual interaction must perforce be negligible. An examination of this bilinear component, in fact, reveals that 79.1 percent of the interaction reached significance, $F(1,7)=8.50$, p. The remainder of the interaction was also significant, $F(8,56)= 2.50$, p.

This bilinear analysis is indicative of the quality of the measurement. However, there are several plausible interpretations with regard to valuation and integration. One solution is to regard one of the two pieces of information as leading to gradual subjective discounting of the truth value of the other, as might be the case with the 'honest thief' in Anderson and Lopes (1974:73). In our manipulation of the original metaphor any novel change, whether a modification of the topic or the vehicle term, makes the entire stimulus combination more artificial, and thus discounts the metalinguistic truth value of the sentence. This might lead to the appearance of a differential valuation strategy or might be related to the use of a conjunctive strategy as a part of an averaging model with differential weighting (see Anderson, 1982).

A conjunctive strategy can be based on the following reasoning: If the politician is not perceived as being verbally adept, it may not make much difference how strong the vehicle term is. The more verbally adept the agent is, the more likely that the vehicle information will be linked with it. Similar logic might apply in the opposite direction when the vehicle provides the focal information to generate a conjunctive strategy, as we suggested earlier in this chapter. A systematic investigation is needed in order to decide whether directionality applies in metaphor valuation.

A different valuation strategy, i.e. differential weighting, was applied here in the comprehension of a novel metaphor. Recalling that comprehension of relatively familiar metaphors was based on orthgonal valuation or simple weighting, we are led to emphasize the role of novelty in determining the valuation strategy in the comprehension process. Yet, additional findings are called for before drawing any conclusion.

Experiment 4
Comprehension of a Metaphor Rooted in American English by Hebrew Native Speakers

In this experiment, as in Experiment 3, we dealt with a metaphoric sentence grounded in the language and culture of American English which was novel for

the Hebrew speaking subjects tested here. It consisted of a concrete nominal topic and an adjectival vehicle. A translation of the prototypical Hebrew sentence was as follows: "An encyclopedia is a gold mine."

Methodological Specifications: Subjects, Materials, and Procedure

In order to allow for possible differences in proficiency, two groups differing in linguistic skills were tested. One group consisted of nine graduate students at Bar-Ilan University ranging in age from 20 to 30, four males and five females. The other group was comprised of ten 18-year-old teacher trainees, all female, performing their National Service at the time of the study. Three subjects were unable to adjust to the measurement conditions and were dropped from the study. Each of the two samples revealed different processing trends in terms of valuation and integration. Thus, their pooled responses are presented separately below.

As in previous experiments, the materials consisted of a metaphoric sentence with four alternative topic terms and four alternative vehicle expressions. The sentence with all stimulus alternatives was as follows: An (encyclopedia/book/dictionary/newspaper) is a (gold/silver/iron/copper) mine. All possible factorial combinations yielded a total of 16 such sentences. Subjects were asked to rate the sentence for the extent to which it implied that the book was valuable. The other details of the procedure were similar to those of previous experiments.

Results and Discussion

Figure 5.4 shows the pooled data for the 9 graduate students.

Vehicle information was relied upon somewhat more than topic information, as implied by the slope of the curves, $F(3,9)= 166.4$ and the distance between them, $F=65.5$ (p). The parallelism of the curves, $F(9,81)$, indicates simple weighting and, as was the case in Experiments 1 and 2 above where judgements were made about noun-noun combinations of topic and vehicle, an averaging rule of integration. This finding, while providing additional support for the validity of functional measurement in the psycholinguistic study of metaphor, raises a question regarding our ability to relate the integration rule used to differences in the semantic nature of the topic information. This time, using a concrete topic and an adjective-noun combination for the vehicle term, we obtained the same rule of integration used previously with an abstract topic and a noun-based vehicle.

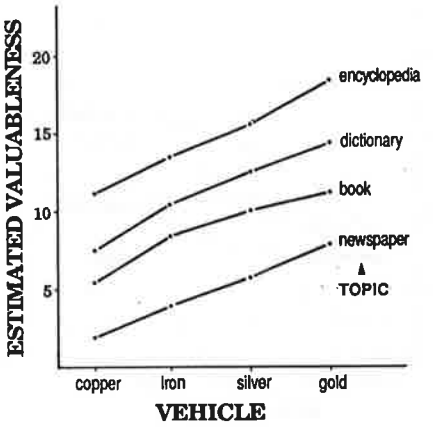


Figure 5.4. Pooled ratings of valuableness as a function of vehicle and topic information in 9 graduate students, all native speakers of Hebrew.

Anderson (1981:2) suggests that the valuation process is capable of taking into account individual differences. Given the psycholinguistic task used here, it is possible that the overall linguistic maturity of the graduate student group tested may have contributed to the parallelism found in the present findings (see discussion below). In that light, we examined a second group from a linguistically less mature population with the same task.

Figure 5.5 depicts the pooled data from the seven teacher trainees.

As was the case previously, vehicle information played a stronger role in the subjects' ratings, as implied by the steep slope of the curves, $F(3,6)=45.8$. For topic, the F-ratio was 5.9, p. The fan shape of the graphic pattern and the interaction term, $F(9,54)=3.3$, p, imply differential weighting. Consideration of the bilinear component in the Topic X Vehicle model indicates that 66.33 percent of the interaction was concentrated there, $F(1,6)=2.19$, 0.05. The residual interaction was not significant, $F(8,48)=1.23$, p0.05, indicating that a trend toward differential weighting or a conjunctive strategy was found among the teacher trainees in this experiment.

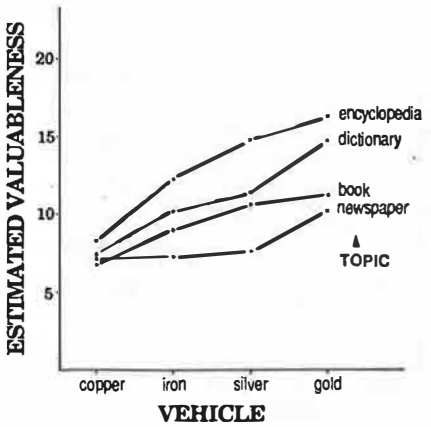


Figure 5.5. Mean ratings of valuableness of the reading material as a function of vehicle and topic information in 7 teacher trainees, all native speakers of Hebrew.

Anderson and Lopes (1974), in discussing judgments of occupational proficiency of a graceful/mathematical/religious dancer/statistical clerk, for example, state that "the occupation noun itself defines the dimension of judgment. Value and weight of the adjective are defined along that dimension, and an inference is requrired to determine them. Clearly the adjective and the noun cannot be integrated as equivalent quantities" (p. 68). The authors presumably studied direct comprehension of how good a person performed his job in response to the adjective-noun stimulus combinations. In metalinguistic judgments of metaphor, adjective-noun combinations seem to conform to the same analysis, i.e. the adjective and noun cannot be integrated as equivalent quantities. It is either the noun which defines the dimension of judgment or, in the case of metaphor, the adjectival element which defines that dimension.

Differential weighting can be interpreted as a reflection of an ad hoc, spontaneous way of assigning importance to the relevant pieces of information embedded in the metaphoric sentence. Put more simply, the interpretive process in this case is assumed to assign importance to the topic information on the background of what is already perceived with regard to the vehicle information, or vice versa.

In this way, anytime the vehicle term (e.g., copper) is degraded in value, the corresponding topic term is degraded accordingly. This can be seen clearly in Figure 5.5 above: The least spread among the four points in the graph depicting topic information is found at the lower left hand corner of the graph where the vehicle information (copper) is degraded. These four points gradually diverge as a function of the valuative strength of the vehicle.

In contrast, parallelism reflects a valuation schema of orthogonality between the relevant sources of information. If we consider ad hoc valuation as reflecting an opportunistic strategy, then a simple weighting strategy can be viewed as motivated by a predetermined rule, independent of trial-based experience. Following the speculation above, our teacher trainee subjects are natural candidates for the application of a differential weighting strategy, which was the finding above. Our comparison graduate student group adopted a predetermined strategy involving simple weighting.

These findings are compatible with Anderson's (1981:2) suggestion regarding the sensitivity of valuation parameters to individual differences. A look back at Experiment 3, however, shows that the college-educated population tended to utilize a differential weighting strategy like the teacher trainees did here. This leaves us as yet unable to decide upon the individual characteristics or the particular linguistic feature governing the valuation strategy. Further systematic empirical purification to unravel possible confounding of personal and linguistic factors may help clarify this issue.

Experiment 5
Comprehension of Metaphors with Bilingual Stimulus Combinations by Hebrew and English Native Speakers

This experiment replicates the semantic and pragmatic elements of the previous study for the metaphor involving valuableness of different sorts of reading material. Here the stimulus material was bilingual, either an English topic appearing with a Hebrew vehicle or a Hebrew topic with an English vehicle.

Methodological Specifications: Subjects, Materials and Procedure

Twenty native speakers of Hebrew and twenty native speakers of English participated. Ten from each group were exposed to sentences with an English topic and a Hebrew vehicle, and ten from each group received stimuli with a Hebrew topic and an English vehicle. All subjects were college-educated, above

the age of 20. One subject in the English native speaker group exposed to stimuli with a Hebrew topic and an English vehicle did not adjust to the measurement conditions of the experiment and was dropped from the study. The entire procedure was the same as reported in the previous experiment.

Results and Discussion

Pooled results for all four groups are presented in Figure 5.6. Three English native speakers, two in the group exposed to an English topic and a Hebrew vehicle, exhibited a different pattern (differential weighting) of results, and thus their data are not included in the pooled findings.

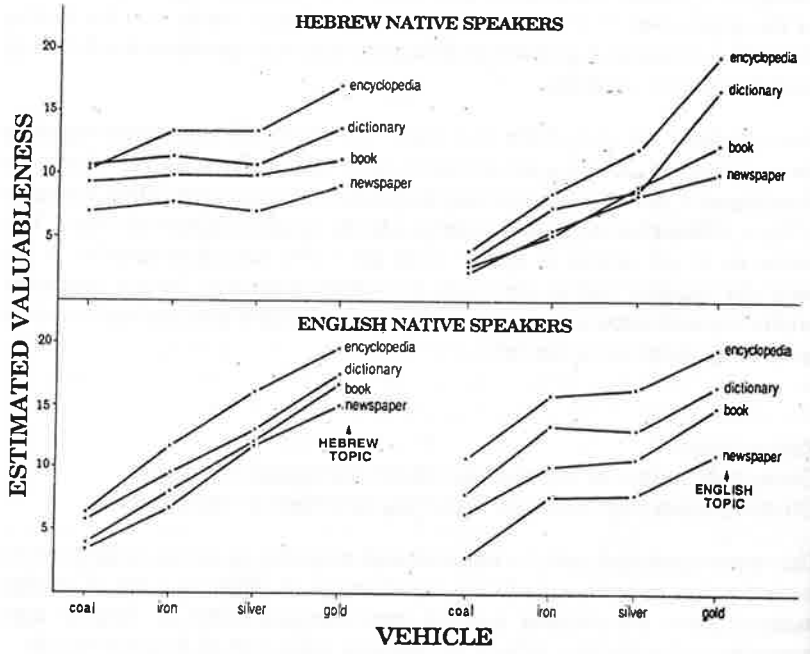


Figure 5.6. Mean ratings of valuableness of the reading material as a function of vehicle and topic information in Hebrew (upper panels) and English native speakers (lower panels) exposed to complementary combinations of English and Hebrew stimuli.

In all four groups, both vehicle and topic were relied upon with the now expected heavier reliance on vehicle in three of the four cases, the exception being Hebrew native speakers exposed to Hebrew vehicles. With regard to orthgonality of valuation, Hebrew native speakers in the two complementary stimulus combinations demonstrated differential weighting, as implied by the fan shape of the graphic pattern in the upper panels of Figure 5.6. Details of the respective inferential statistics are summarized in Table 5.1.

		Vehicle		Topic		Vehicle X Topic	
		df	F	df	F	df	F
Heb Nat Speakers	Heb Top/ Eng Veh	3,9	25.46	3,9	3.44	9,81	2.51
	Eng Top/ Heb Veh	3,9	9.56	3,9	13.46	9,81	5.65
Eng Nat Speakers	Heb Top/ Eng Veh	3,7	127.35	3,7	14.01	7,72	1.99
	Eng Top/ Heb Veh	3,7	47.87	3,7	21.94	7,72	0.44

Table 5.1. Summary ANOVA statistics for vehicle and topic main effects and interactions for judgments of valuableness of the reading material in Hebrew and English native speakers exposed to complementary combinations of English and Hebrew stimuli.

All main effects in Table 5.1 confirm the visual impression of the graphic display in Figure 5.6, i.e. they were significant in all four groups, and vehicle effects were stronger in three of the four. The impression with regard to orthogonality of valuation finds support in the interaction terms as well. The interaction terms for the Hebrew native speakers exposed to Hebrew and English vehicles were significant, F(9,81)=5.65 and 2.51, respectively (p). An examination the Topic X Vehicle model shows that for the Hebrew and English vehicles, respectively, 62.45 and 73.23 percent of the interaction was related to the bilinear component, F(1,9)=1.57 (p0.05) and 6.59 (p). The residual interaction did not reach significance in either case, F(8,72)=1.05 and 1.49 (p0.05), pointing to a differential

weighting or conjunctive strategy on the part of Hebrew native speakers in the two stimulus conditions.

With regard to the two groups of English native speakers the visible graphic parallelism is reflected in non-significant interaction terms, $F(9,63)=1.25$ for the Hebrew vehicle condition and $F(8,72)=1.99$ for the English vehicle condition ($p0.05$), implying a simple weighting strategy.

Beyond differences in stimulus combinations, Hebrew native speakers applied a strategy of differential weighting in the metalinguistic comprehension of metaphor. The fact that the gold mine metaphor is rooted in the American English vernacular might be related to the consistent language-related differences. For English native speakers this metaphor is well-ingrained and accomodated into their usage patterns. It is plausible to transform the language of simple weighting or orthogonal valuation into the more mundane terminology of an a priori subjective valuation of each source of information. Such a strategy is reflected in our case in the constant weight assigned to all levels of vehicle and topic information.

An English native speaker presumably has pragmatic expectations whenever he is primed with even a slight signal belonging to the familiar metaphoric pattern which is assumed to serve as a peg for retrieving the entire complex of this specific linguistic gestalt. In this way he is supposed to operate in a top-down mode; the weighting of a specific source of information applies to all possible substitutes of the original elements of the familiar metaphor, whether vehicle or topic, as long as they are not too far from the source material. This reasoning accounts for the simple weighting applied by the English native speakers in the valuation of a familiar metaphor.

The Hebrew native speaker, following the above reasoning, is assumed to have adopted an ad hoc, bottom-up way of coping with the task of comprehending uncertain (SL) metaphoric material. It implies that, not having a good grasp of the material, the subject re-evaluates any stimulus combination, degrading one source as the other is perceived to be less sensible. If this assumption is valid, then establishing conditions which are lexically and pragmatically more integrative (as compared to the relatively artifical stimulus combination used in this experiment) should lead to the use of an a priori, top-down strategy.

Experiment 6
A Further Test of an A Priori Strategy in Comprehending an English Metaphor by English Native Speakers

In order to generate more lexically-unified sentence components, we constructed a metaphoric expression where topic and vehicle belong to a strongly associated lexical network drawn from the context of colloquial speech. This afforded the language user with more familiar and natural stimulus material. The target sentence read: 'The opera singer had a creamy butter voice.' To provide even more natural flavor to the comprehension task, subjects were instructed this time to make ratings of a direct comprehension question, i.e. how pleasant the voice sounded.

Methodological Specifications: Subjects, Materials and Procedure

Eight native English speakers, all university graduates above the age of 25, participated. The materials consisted of the following topic terms: opera singer, adolescent boy, and heavy smoker. They were factorially combined with the following vehicle terms: creamy butter, nutty chocolate, and dry, burned toast voice. The task and procedure were similar to those used in previous experiments.

Results and Discussion

Pooled data for the eight subjects are presented in Figure 5.7.

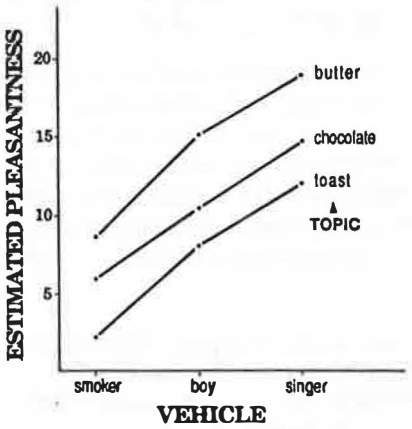


Figure 5.7. Mean ratings of voice pleasantness as a function of vehicle and topic information in 8 native speakers of English.

An inspection of Figure 5.7 indicates that both vehicle and topic terms were found meaningful in the comprehension of the target metaphor, as can be seen by the slope of the curves and the well-spaced distance between them, $F(2,7)=47.77$ and 35.96, respectively (p). The insignificant interaction term, $F(4,28)=2.24$, p0.05, confirms the apparent parallelism in the graphic pattern. These findings provide more generality to our conclusion from the previous experiment that lexical material which is readily integrable lends itself to top-down valuation in the form of simple weighting.

These findings may pave the way for a working hypothesis based on the notion that whether a top-down or bottom-up mode of processing is expected depends on the lexical/pragmatic cohesion of the metaphoric sentence. A metaphoric sentence with topic and vehicle phrases strongly associated with each other would prime or touch off simple weighting based on a priori or orthogonal assignment of importance to each of the relevant dimensions of information. If, however, topic and vehicle do not have a strong lexical interconnectedness, differential weighting which reflects a bottom-up or conjunctive mode of processing, is likely to be primed.

CHAPTER SIX: Invariance in the Integration of Primary and Second Language Lexical Hints

The present study shifts the focus of our search for integrative SL processing to the lexical level of analysis. In the previous chapter, topic and vehicle information were manipulated at the sentence level of processing. The present study further constrains the comprehension process, examining metalinguistic judgments involving combinations of hints from a speaker's primary language and secondary language. Another parameter investigated in this study distinguishes between individuals of different levels of proficiency in SL use.

Experiment 1
Metalinguistic Judgments of Heterogeneous and Homogeneous Hinting Information in Speakers of English as a Second Language

Methodological Specifications: Subjects, Materials, and Procedure

All subjects were female undergraduates ranging in age from 18 to 25. Two groups of native Hebrew (PL) speakers, one proficient and one non-proficient, equated for relevant variables except English proficiency, participated in this experiment. Each group consisted of 24 students. Both groups were equally divided into three sub-groups, each of which was exposed to one of three different combinations of PL and SL hint words in a comprehension task described below.

The materials for this experiment consisted of the following sentence: "A medlar is like an apple." (A medlar is a fruit about the size of a fig with a slightly sour taste.) The target lexical item, 'medlar,' was determined to be unfamiliar by a pretest in which subjects were asked to explain the meaning of the target sentence. With regard to the term 'apple,' there is no reason to expect less than complete familiarity as far as its lexical and functional properties are concerned. The term 'medlar,' due to the predicate 'like' which follows, gains a certain degree of familiarity through its association with apple. At the same time, the term 'like' implies only partial identity between medlar and apple. Thus, some degree of uncertainty remains. Food-like properties can be assumed to establish the similarity relation. In this way, the comprehension process is fully determined

by a combination of known and unknown lexical items linked with each other by a term predicating partial equivalence.

The language user was exposed to different combinations of hint items varied along a generalized dimension of semantic distance from the fruitlike nature of a medlar. The helping information was always presented as a combination of lexical items. The subject was instructed to relate to both pieces of information as relevant. Presentation of one fruit as a hint item is expected to prime a working hypothesis on the part of the subject that 'fruitiness' is the relevant dimension on which to anchor the judgment. But it is the other hint included in this combination which will determine how much redundancy there is with regard to the crucial question, i.e.

To what extent is fruitiness a relevant dimension? If the second hint is another fruit, considerable redundancy emerges, since both hints (which are assumed to be relevant) fit the dimension of fruitiness. If, however, the other hint is an abstract term, the preliminary hypothesis that fruitiness is what really links medlar with apple cannot be fully supported. If both hint words are abstract, a large degree of confusion should result due to the distance between the fruitiness of the word 'apple' and the abstractness of the hinting information.

An across-language unity is assumed to characterize this comprehension process. PL and SL information are assumed to play an identical role in such hypothesis testing. A hinting combination comprised of a name of a fruit in PL and a name of another fruit in SL is assumed to reduce the same amount of uncertainty as does a combination of names of two fruits in SL. This leads to an overall hypothesis of across language invariance with regard to the specific process probed here, i.e. reduction of uncertainty about relevant dimensions for lexical hypothesis testing.

A total of sixteen cards was presented, each one containing two hint words. The two hint words belonged to one of three conditions, one heterogeneous and two homogeneous. For the heterogeneous condition one word was in English and one in Hebrew. For the homogeneous conditions both hint words were presented in the same language, involving a combination of two Hebrew (PL) words and a combination of two English (SL) words.

The English and Hebrew hint words were comprised of four pairs; each pair belonged to the same semantic category. None of the pairs were translation equivalents. Each of the three hint conditions mentioned above involved a complete factorial combination (4 X 4) of two sets of hint words (English/

English, English/Hebrew, and Hebrew/Hebrew). Thus, the complete design consisted of 16 combinations of two hint words, each pair appearing on a different card.

For the heterogeneous condition, the English hint words consisted of the following four English lexical items: peach, bread, cat, and love combined with the following Hebrew words: 'agas' (pear), 'uga' (cake), 'kelev' (dog), and 'simxa' (happiness). For the English-English homogeneous condition, the same four English words were factorially combined with the following four lexical items: pear, cake, dog, and happiness. For the Hebrew-Hebrew condition, the stimuli were: 'agas' (pear), 'uga' (cake), 'kelev' (dog), and 'simxa' (happiness) combined with: 'afarsek' (peach), 'lehem' (bread), 'hatul' (cat), and 'ahava' (love).

All subjects were individually tested in a separate room on the university campus. The experimenter presented the subjects with the stimulus sentence ("A medlar is like an apple.") All subjects reported that they did not know the meaning of the word 'medlar.' Following the introduction of this sentence, each subject was presented sequentially with the sixteen pairs of hint words. The subject was then asked to evaluate the degree to which the words on the cards helped her to understand the meaning of the target word. The evaluation was made on a 20-point graphic rating scale.

Subjects were asked to consider meanings only and to ignore associations and/or similarities of shape, color, etc. As a warm-up routine, each subject was given five or six practice pairs to rate. Their judgments on these practice pairs were discussed to ensure understanding of the task. The session lasted approximately half an hour. The order of presentation of the sixteen cards was determined by shuffling the cards prior to each session.

Findings

The pooled findings of the experiment are graphically depicted in Figure 6.1.

As implied by the lower-left to upper-right slope of the curves and their spacing, all hints in all conditions were valuated as helpful in understanding the target item. Main effects for the English-only hints (English/English) were significant for the students who were proficient in English (see panel 1). The two respective F-values (df=3,7) were 61.00 and 121.71. For the non-proficient students presented with English-English hint combinations (panel 4), F=10.16 and 17.64.

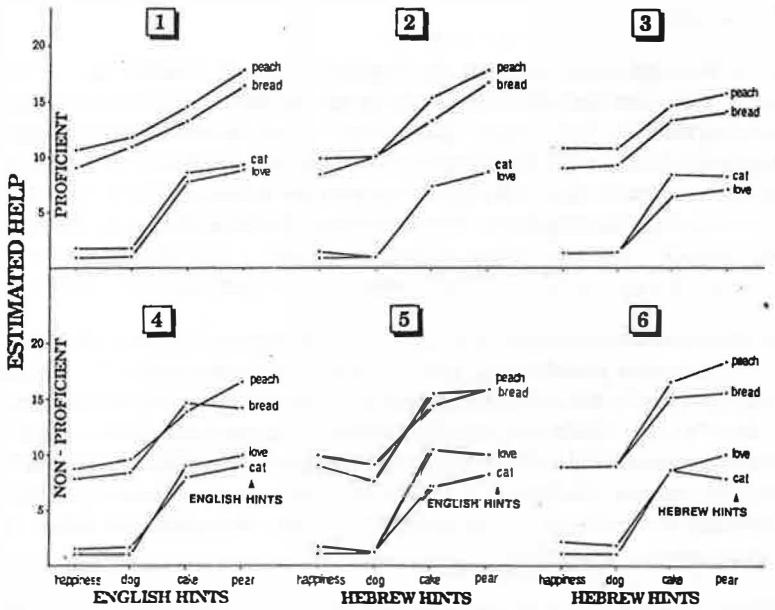


Figure 6.1. Mean estimated help under conditions of heterogeneous and homogeneous hint words for proficient (upper panels) and non-proficient (lower panels) speakers of English as a second language.

Similar findings emerged in the case of simultaneous presentation of Hebrew and English hints, F=68.40 and 32.46, respectively, for the proficient students (panel 2) as well as for the non-proficient ones (panel 5), for which the respective F values were 12.74 and 12.41.

The same pattern holds for simultaneous presentation of two Hebrew hints for both proficient students (panel 3), F=65.07 and 52.06, and for non-proficient students (panel 6), F=15.23 and 26.10. All 12 p-values were far below the 0.01 alpha criterion.

All six configurations in Figure 6.1 exhibit parallelism among the curves. The insignificance of five of the six interaction terms provides inferential support for this. The F(9,63) scores were 1.04, 1.43, 1.90, , , and 2.19, for conditions 1-6, respectively.

In a general sense, the unity of the overall pattern of these findings is striking, and it may well be concluded that both SL and PL hints were relied upon to relatively the same extent. Moreover, these results were replicated in both homogeneous and heterogeneous combinations of hints and in proficient as well as non-proficient SL users. It might be worth noting that the same integrative pattern of processing appeared even when the main effects were considerably weaker (i.e., in the non-proficient subjects). Another invariant trend can be seen in the bipolar convergence of the curves as well as in their two-tiered shape. In several cases there was even a complete overlap of the curves (see panels 2, 3, and 6) and a lack of difference within either the two lower or two upper points of the vertical dimension. This indicates a tendency to classify the hints into two general categories, one close to 'apple,' i.e. food, and one distant from it.

Experiment 2 examines whether reliance on multi-source (in our case two) hints and an invariant integrative pattern of processing can be found in an inverse linguistic sample, namely, native English speakers who use Hebrew as a SL.

Experiment 2
Metalinguistic Judgments of Heterogeneous and Homogeneous Hinting Information in Speakers of Hebrew as a Second Language

Two groups (proficient/non-proficient) of native Hebrew speakers for whom English was a second language participated in this experiment. Each group consisted of 24 students in a design comparable to those in Experiment 1. Again, each group was equally divided into three sub-groups. Subjects were individually exposed to the same three combinations of PL and SL hint words and a comprehension task identical to that used in Experiment 1. The materials, the stimuli and the procedure were also identical to those used in Experiment 1 with one exception: A Hebrew sentence, which was an exact translation of the English target sentence used in Experiment 1, was presented for comprehension.

Figure 6.2 depicts the pooled findings of Experiment 2.

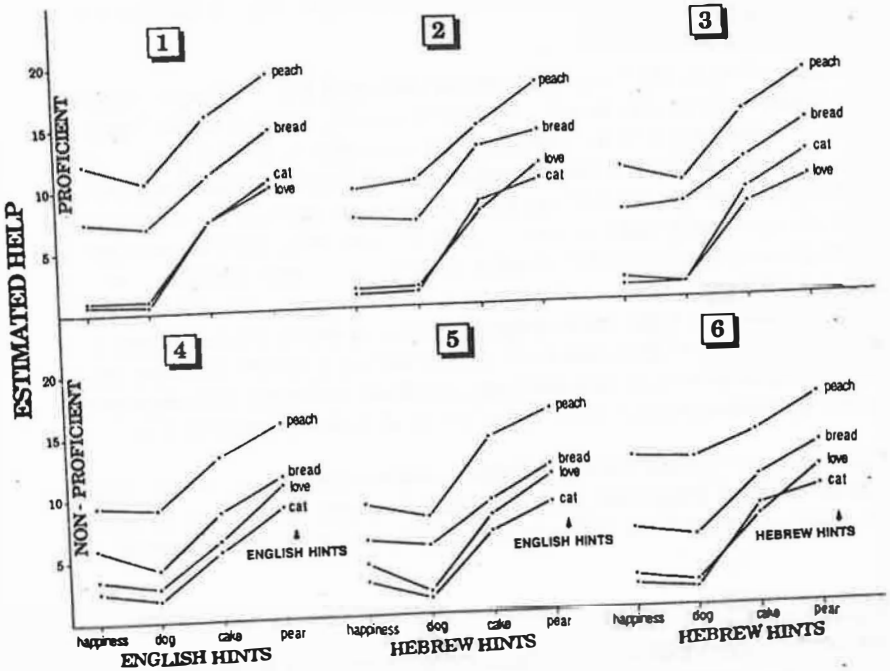


Figure 6.2. Mean estimated help under conditions of heterogeneous and homogeneous hint words for proficient (upper panels) and non-proficient (lower panels) speakers of Hebrew as a second language.

It is evident from all six panels in Figure 6.2 that subjects' judgments of the degree to which the words on the cards helped in understanding the meaning of the target word indicated reliance on both hints. This was the case whether the hint words were English/English (F values were 495.59/440.38 for the proficient subjects and 128.11/91.22 for the non-proficient subjects), English/Hebrew (F values were 298.50/169.18 and 73.50/44.03, respectively), or Hebrew/Hebrew (171.58/443.85 and 219.11/112.50). All 12 p-values were far below the 0.01 criterion.

As in Figure 6.1 (Experiment 1), the configurations in Figure 6.2 demonstrate parallelism among the curves. The F(9,63) ratios for the interaction terms were 7.16, 1.69, 6.65, , 1.76, and 1.15. for conditions 1-6, respectively. Even though two out of the six interaction terms attained statistical significance (the first and the third), the overall salient visual parallelism among the curves suggests that the significance of these two interaction terms is, rather, a by-product of incidentally large inter-individual variance.

These findings are similar to those of Experiment 1 and support the conclusions drawn on the basis of that experiment: Both SL and PL hints were relied upon by users of two languages, English and Hebrew, regardless of which language serves as PL or SL, whether the combination of hints was heterogeneous or homogeneous, and regardless of whether the language user was proficient or non-proficient in the SL. The bipolarity observed in the judgments of native English speakers (in Experiment 1), is not replicated here, however. A more differentiated trend is evident in the judgments of native Hebrew speakers (this experiment) as compared to the native English speakers. This can be seen in the tripolar spread of the responses to both sources of information and the merging of only the two lower levels (see Figure 6.2). This difference between native English and native Hebrew speakers is consistent across all conditions within each sample.

While quite consistent evidence has been presented, the use of a single response measure does not yet permit a satisfactory generalization of the findings, especially since the comprehension task employed was indirect. In order to clarify this point, Experiment 3 replicates a portion of the previous design while instructing the subjects to perform direct comprehension of an unfamiliar SL lexical item.

Experiment 3
Likelihood Judgments of Heterogeneous Hinting Information

All eight subjects in this experiment were native speakers of English who had been in Israel for over three years and had passed the Hebrew examination requirements for an undergraduate degree at Bar-Ilan University. Their proficiency in both languages was that of native speakers. Thus, they may be considered to be fluent bilinguals.

The stimuli were constructed from an English target sentence and pairs of hint words identical to those used in Experiment 1. One hint word was always English

and the other Hebrew (English/Hebrew hint combination). The subject was instructed to use these hints in order to evaluate how likely it is for a medlar to be like an apple. The rest of the procedure was identical to that used in the previous experiments.

The pooled results of the direct compehension experiment are presented in Figure 6.3.

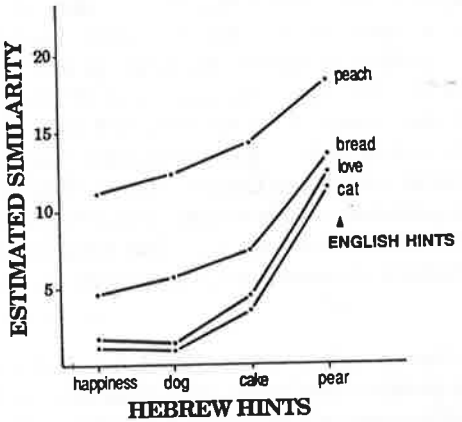


Figure 6.3. Mean estimated similarity under conditions of heterogeneous and homogeneous hint words for native speakers of English.

The graphic structure in Figure 6.3 is generally similar to those in the previous figures in this chapter. Both English and Hebrew hints contributed to direct comprehension, $F(3,7)=134.97$ and 166.79, respectively. The two p-values were again far below the 0.01 criterion.

The insignificance of the interaction term, $F(9,63)=1.54$, again implies parallelism. In contrast to the orthogonality implied by this parallelism, the visual convergence of the curves may point to a latent trend of differential weighting. This would suggest that more informative hints from one language may lead to less reliance on hints from the other language. To decide between this impression

and the above formal statistical inference, a finer experimental analysis is recommended.

Experiment 4
Linguistic Examination of Idiom Comprehension: Connotative Distance and Lexical Content

Idioms take on a special role in second language use. They are almost entirely culture-dependent and thus actually language dependent. A talented SL user can easily acquire a functional set of grammatical rules as well as lexical networks and yet have difficulties with idiom use, perhaps due to the limited potential for transfer from a PL repertoire of idioms. It is well agreed upon among second language teachers that mastery of figurative language is a critical sign for fluency in a foreign language. This chapter explores a possibility for functional examination of SL idiom comprehension. Idioms are sampled here from the larger domain of figurative language, the other most commonly used figurative form, metaphor, already having been dealt with (see Chapter 5).

Fifty four subjects, 37 proficient and 17 non-proficient, were drawn from Bar-Ilan Unversity's English as a Foreign Language Program. Their proficiency was determined by a national placement examination. The design involved presentation of an unfamiliar American English idiom, 'be off.' As a screening test to determine prior familiarity with the idiom, subjects were asked to translate the target phrase into their native Hebrew. None of them was able to correctly identify the meaning of the idiom.

In addition to the target idiom, stimulus materials consisted of pairs of hint items from Hebrew (PL) and English (SL). Both stimulus dimensions had two levels, one very similar in meaning to the target idiom and the other less similar. Each target idiom was paired with a combination of two hint items, one Hebrew and one English, which were presented on cards one at a time. The subject's task was to estimate the amount of help he or she felt the pair gave to an understanding of the idiom.

The entire arrangement consisted of four presentations of combinations of the idiom and hint pairs, forming a complete 2 X 2 factorial design. This design served as a basis for manipulation of the connotative distance between the Hebrew and English hint items. Two dichotomous manipulations were factorially combined to form the following four versions (2 X 2) of the original design.

One manipulation was synonomy/translation. A second consisted of two different lexical contents. The full set of 16 stimulus pairs were as follows:

Synonymy, Content A: exit, flee; la'azov (leave), livroax (escape)
Synonymy, Content B: leave, escape; lacet (exit), l'himalet (flee)
Translation, Content A: exit, flee; lacet (exit), l'himalet (flee)
Translation, Content B: leave, escape; la'azov (leave), livroax (escape)

These items were counterbalanced within and across the four manipulations.

Subjects were tested in their classrooms, first being instructed in the use of rating scale procedures. A graphic rating scale ranging from 1 to 20 with markings only at the endpoints was used. Subjects were told to mark an X at the point on the scale indicating the amount of help they felt the pair of hint items gave in understanding the idiom 'be off.'

Figure 6.4 presents graphs for the experiments included in the overall design.

In all eight graphs in Figure 6.4, a strong main effect for both English and Hebrew hints is evident, as portrayed by the slope of the curves and the distance between them, respectively. The respective inferential statistics displayed in Table 6.1 corroborate this finding.

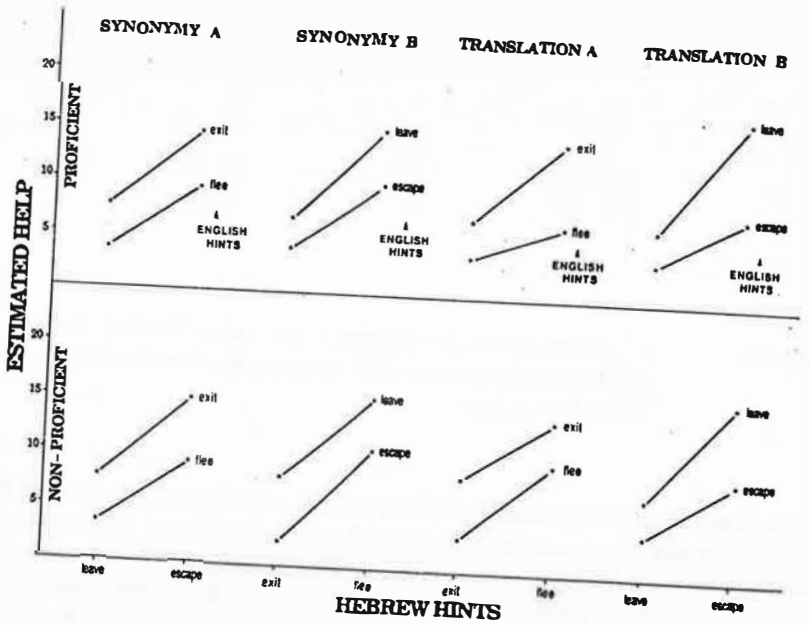


Figure 6.4. Mean estimated help of English and Hebrew hints in four experiments varied by synonymity and lexical content for two levels of proficiency (proficient, n=37; non-proficient, n=17).

	Synonymy A			Synonymy B			Translation A			Translation B		
	Eng	Heb	ExH	Eng	Heb	ExH	Eng	Heb	ExH	Eng	Heb	ExH
High	131.8	177.5	2.4	57.1	279.4	3.6	93.1	198.6	6.8	193.8	203.8	18.5
Prof			ns			ns			*			***
Low	63.1	79.0	1.0	38.0	11.7	< 1	23.1	68.8	2.8	61.4	60.5	11.8
Prof			ns			ns			ns			**

Table 6.1. F-scores for the complete factorial design, English X Hebrew hints, in four experiments varied by synonymity and lexical content for two levels of proficiency

There is a gradual change shift from parallelism in the left panels of Figure 6.4 to a left-to-right divergence between the lines in the right panels. This trend is more explicit in the responses of the high proficiency subjects. The F-ratios in Table 6.1 representing the interaction coefficients corroborate the impression regarding this pattern. For low proficiency subjects, the F-ratios for synonymous hint items (the two leftmost interaction terms) are practically nil, while the two rightmost terms, reflecting translation equivalents highly similar to the target idiom, indicate some tendency for left-to-right divergence of the curves, as can be seen in the respective panels of Figure 6.4. The rightmost coefficient is significant at the 0.01 level. For the high proficiency subjects this is even more pronounced, both interaction coefficients for the translation versions being significant (the two rightmost scores), supporting the visual impression of left-to-right divergence of the curves in the two upper rightmost panels of Figure 6.4

Parallelism presumably represents simple-orthogonal weighting of the two hint items, while fan shaped curves can be said to indicate differential-interrelated weighting. The latter strategy happens to be applied here in the use of translated hint items, which are considered closer in meaning to the target idiom than the synonymous hint items. While this finding has intuitive appeal and can be accepted as a basis for a working hypothesis in future experimentation, its conceptual nature is yet to be determined.

Experiment 5
Psycholinguistic Examination of Idiom Comprehension

In order to challenge the generality of the above findings under differing extra-linguistic conditions, an abbreviated version of the above design was implemented. This included only the manipulation of connotative distance, which appears to be provocative for integration processes. Each subject was thus exposed to two English X Hebrew (2 X 2) factorial experiments, one with pairs of synonymous hints and the other with translation equivalents. These experiments were conducted sequentially in a single session, and the order of presentation was counterbalanced.

The entire design was applied to a group of 14 native Hebrew speakers, fluent in English as second language. All were students majoring in English at Bar-Ilan University. Subjects were individually primed with 10 minutes of English-English lexical definitions drawn from the American College Dictionary (Barnhart, 1968). They were provided a list of 12 motion verbs and asked to look up their definitions. Following the look-up procedure, subjects were allowed the balance of time for study of the words and definitions. Being in a state of mind prompted by extensive use of unilingual definitions is presumed to constrain comprehension in a way that discourages bilingual processing.

The results are displayed in Figure 6.5

The slope of the curves as well as the distance between them in the two panels of Figure 6.5 indicate that both English and Hebrew hints were regarded as helpful to the comprehension of the idiom 'be off.' The curves in the right panel, representing the translation equivalents, seem to follow a clear, left-to-right divergent pattern, while those on the left are close to being parallel, with a little left-to-right divergence.

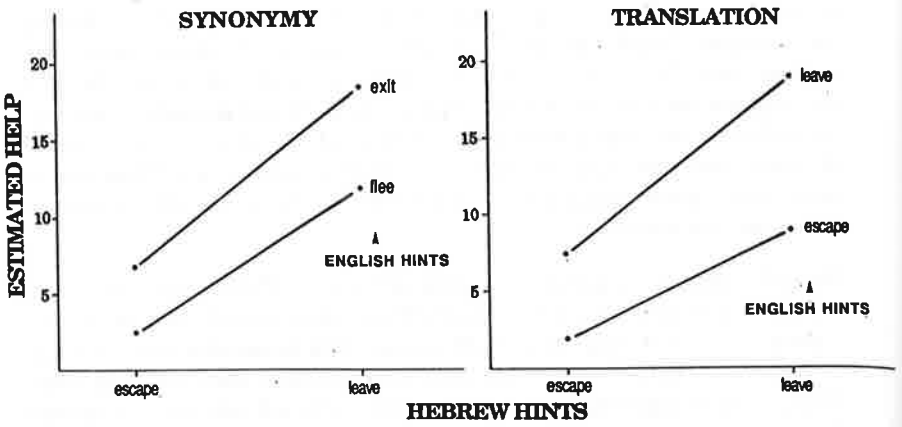


Figure 6.5. Mean estimated help of English and Hebrew hints in two experiments (synonymy and translation) in 14 fluent EFL users exposed to unilingual lexical priming.

The inferential statistics support the conclusion with regard to significance of main effects (df=1,13; p) both for English (F=161.48) and Hebrew (241.74). The interaction term (7.94) was significant (at the 0.05 level). This supports a conclusion regarding the tendency for differential weighting under conditions of unilingual priming.

For the left panel, representing judgments of synonymous hints, main effects (p) for English (F=19.16) and Hebrew (109.19) were significant. The interaction term (1.79), however, did not attain significance, thus indicating a tendency for simple weighting. Overall, these findings replicate those reported in the previous experiment, showing meaningful differences in the weighting strategy in accord with the connotative distance between the hint words.

In an attempt to replicate this invariant trend, 13 fluent EFL students were primed with English-Hebrew definitions taken from the Complete English-Hebrew

Dictionary (Alcalay, 1970). All other procedures were the same as for the EnglishEnglish priming experiment.

The findings are graphically portrayed in Figure 6.6.

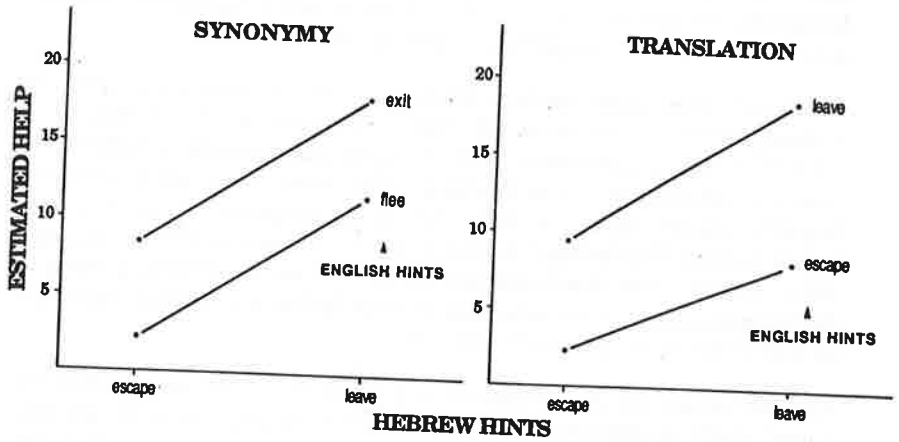


Figure 6.6. Mean estimated help of English and Hebrew hints in two experiments (synonymy and translation) in 13 fluent EFL users exposed to bilingual lexical priming.

All relevant graphic characteristics of both panels in Figure 6.6 indicate that valuation and integration typicalities of estimating help from hinting combinations under the influence of bilingual lexical priming are similar to those typicalites found following unilingual lexical priming. All main effects in both the synonymy and translation experiments were significant at the 0.0001 level. The left-to-right graphic divergence in the right panel, which represents the translation stimuli, is supported by a significant interaction term (F=5.28, p), thus pointing to differential weighting of the two sources of hinting information. The nearly parallel curves in the left panel represent simple weighting due to the negligible interaction coefficient (F<1).

Conclusions

It was found here that stimulus combinations, whether homogeneous (two English hints or two Hebrew hints) or heterogeneous, were consistently processed using the same valuation strategy. This trend was found among language users of entirely different proficiencies and from two different native languages (either Hebrew or English as PL). The applicability and possible contribution of Information Integration Theory to research on SL use is implied by consistent utilization of the same comprehension patterns reported above.

Functional measurement has been shown to be a useful means to account for various psycholinguistic issues (cf., Oden and Anderson, 1974; Oden, 1977, 1978, 1986). Those applications, however, did not purport to deal with secondary linguistic sets such as second languages. The present study approached two linguistic sets (PL and SL) as molar units in the comprehension of unfamiliar lexical material. The finding that different lexical instances of the two languages were combined during the comprehension process can be viewed as pointing to the integrative nature of the process in question. It does not, however, shed light on how details of this process fit together.

A theoretical account of a mechanism responsible for the process in question is a need which yet remains. Research into intregative processes in SL use may benefit from an examination of proposals about how different cognitive units combine with each other during the process. Such work would complement the focus here on examination of valuation strategies and integration rules which are probed by metalinguistic and more direct comprehension tasks.

With regard to the well-accepted distinction between primary and secondary language, our findings seem to place an empirical barrier to the meta-theoretical assumption that SL should be treated as a legitimate independent psycholinguistic entity. The findings consistently support an invariant pattern of information integration across different linguistic stimulus compounds that were based on the differentiation between PL and SL. The results should be considered, then, to weaken the argument that SL is a unique psycholinguistic reality. Rather, PL and SL can be seen as two poles of the same generalized linguistic universe.

In general the experiments reported on in this chapter seem beneficial for the attempt to identify both stimulus and response modes where integration of PL and SL information might play a meaningful role. Several working hypotheses for future research can be inferred from this work: a. If an SL user is provided with combinations of bilingual hint words and required to make a metalinguistic

response, integrative processes emerge. b. Hint words from both PL and SL are subjectively weighed as meaningful (with some preference for PL information). c. A linguistic parameter such as connotative distance (between PL and SL hint words) seems to constrain integration strategies such as those involving simple weighting and differential weighting. d. These trends were found to be invariant with regard to a psycholinguistic factor which can be conceived as a 'lexical state of mind' and operationalized as priming with the use of either unilingual or bilingual lexical definitions.

PART III: SUBSTANTIVE AND METHODOLOGICAL CONSIDERATIONS

CHAPTER SEVEN: Concluding Remarks

The empirical chapters provide limited yet encouraging indication that integrative processing is prompted where the SL user is challenged with SL material of partial certainty. This was observed in several linguistic contexts. We began our search (in Chapter 4) by testing the viability of the integration idea at the narrative level, only indirectly addressing SL issues. It was found that non-proficient subjects performed the same as fluent SL users, integrating structural and semantic information in a metalinguistic task as well as in direct comprehension of narratives in English as a SL. Another exemplification of within-SL integrative processing was observed (in Chapter 5) within the boundaries of a smaller discourse domain, i.e. the sentence level. Hebrew and English native speakers were found to integrate topic and vehicle information in similar ways across PL and SL in comprehending SL metaphors.

In the comprehension of an unfamiliar lexical item embedded in a sentence context (Chapter 6), more essential evidence was found, this time for integration of PL and SL information in SL comprehension. This issue was further examined in a study of the comprehension of an unfamiliar idiom (Chapter 6) where that idiom was isolated from its customary linguistic context. More evidence for the integrative role of PL and SL information was found when the same paradigm served as a basis for manipulation of priming with unilingual and bilingual lexical definitions.

7.1 Use of PL Information in SL Comprehension

Semantic and pragmatic mastery of a second language is acquired by anchoring new linguistic units in other linguistic schemata already acquired by the user. There are two ways in which this can occur. One approach relates to language acquisition as if it has a sterile nature and assumes that an attempt to associate SL material with PL knowledge may contaminate the former. Accordingly, it would be desirable for new SL material to be superimposed on already acquired SL material, rather than reliance on material from one's PL.

Many foreign language teachers seem to adopt an approach consistent with this thinking and promote it to a degree that gives it the aura of a myth. Following their line, a foreign or second language learner should 'suffer' in order to purify the process of acquiring new linguistic structures. The most salient implication of this approach is the instructional imperative to use only a unilingual dictionary. A SL student who, on the sly, uses a bilingual dictionary does so with the belief that he will pay a price for this practice at some point in the future.

A second approach assumes that during second language processing a person cannot ignore linguistic knowledge in PL, which is continually operative before and after SL use. In this light, Schachter's (1974) finding that Japanese learners of English avoid use of relative clauses in their second language can be accounted for by the assumption that avoidance of constructs in SL may be due to an inability to fluently transfer parallel constructs from PL. Relative clauses in Japanese present such a problem in the learning of English. Trying to avoid the whole linguistic system which a language user has acquired leaves him without an operative framework for language processing. A language user who follows the didactic imperative of the previous 'SL-only' approach might threaten his operative linguistic framework. In doing so, he gives up his own basis without yet having another one. Associating an unfamiliar SL term or expression with a similar PL linguistic unit can provide a reliable channel for processing new material. A daily application of this approach is the use of a bilingual dictionary in SL comprehension tasks.

Common to these two approaches is the integrative nature of the comprehension process. The unilingual approach assumes homogeneous, within-language integration, while the bilingual approach entails more efficient SL processing via heterogeneous, across-language integration. To decide empirically between these two alternatives, experimental competition is called for. Our attempt to prime unilingual versus bilingual integrative processing by English-English and English-Hebrew dictionaries respectively (see Chapter 6) can only serve as an exploratory beginning for operationalization of this issue. Lack of meaningful differences in comprehension responses following different modes of dictionary use do not imply failure to reject the null hypothesis. There is no reason as yet to doubt the priming effect. The widespread use of both bilingual and unilingual dictionaries attests to this. It is the nature of the response to lexical priming which may deserve more careful consideration.

Evaluative, or metalinguistic judgments, as used here, might be relatively insensitive to the specific state of mind induced by lexical priming. It may be

beneficial to follow the experience of priming by direct comprehension tasks which allow the language user a choice between open, divergent comprehension and more circular, analogous processing. Production tasks allow considerable volition in terms of modes of language use. Combining them with different psycholinguistic states of mind (such those generated by unilingual and bilingual priming) may help the desirable phenomenon to reveal itself.

7.2 Friction Points Between PL and SL

The work and ideas presented in this volume have dealt with one prinicpal thesis, i.e. languages in use are related to each other in the process of SL comprehension. It is assumed that cognitive entities belonging to different languages a person has acquired are not partitioned or isolated in the SL user's mind and are not processed independently. These relations are conceived as a valuative, integrative, and functional whole.

We advocated the use of bilingual dictionaries in preference to unilingual ones. It is too early, however, to merge this already formulated, yet immature idea with other, more developed psychological models which deal with information processing. We chose to affiliate our idea with an approach that examines fundamentals of SL use, such as its functional, multifaceted, and integrative nature, in the context of a substantive theory in cognitive and general psychology, IIT.

A first step to be taken is to derive molar and molecular units which may play a role in the overall process of SL use. The most general classification of molar units is the division between PL and SL information. Another molar classification is PL/SL syntax, PL/SL semantics, etc.

Semantic nodes might serve as entries for identification of molecular units. In the lexical domain, friction points can be observed when a speaker embeds a noun from his PL in a second language sentence. This presumably arises from an incomplete lexical network in SL. Frequently, a second language speaker does this linguistic jigsaw without paying attention to the switch. The spontaneity of this sort of communicative solution for lexical friction can be conceived as ad hoc support for our integrative approach. An alternative result of such a bilingual dilemma would be blockage of the flow of processing. In future experimentation it would be advisable to design systematic confrontation between these two alternative hypotheses. The focus would be on the use of friction points as a testing ground for a debate between the integrative approach (which we advocate here) and the 'purist,' same-language approach, as found in Krashen and those

who advocate the use of unilingual dictionaries and audio-lingual teaching methods (see Brown, 1987).

Another source of friction points spans both phonological and semantic domains. The SL speaker slips in the middle of a sentence into his PL, never returning to the SL. This friction can occur between two semantically identical words, phrases, or clauses, one from SL and one from PL. In this case only one of the two competing words (i.e., a lexical item from one language only) can fill the semantic vacuum within the sentence. The PL choice is the one that usually takes priority. Unlike the previous case involving lexical friction, here the speaker does not suffer from an incomplete lexicon. Rather, he has faster access to the compatible PL node. The identity of meaning is what generates friction here. Recalling that with the previous sort of friction the switch involves only one word, the question emerges why in this case the switch is for the remainder of the sentence.

The first case we mentioned above is similar to what Poplack and Sankoff (1987) call borrowing and what McLure (1977) has called code-mixing. Embedding a PL noun in SL discourse involves friction only at the lexical level. It is a substitution of one lexical item for another. The second case is similar to what McLure has called code-changing and what has been more widely referred to in the literature as code-switching. This implies a change in language as well as in syntax, semantics, etc.

Our reference to friction points is restricted to those resulting from code-switching in one direction only, viz. from SL to PL. This form of code-switching can be distinguished from PL-to-SL shifting, which has a more sociolinguistic orientation, being used by the speaker for focus, emphasis, or to impress the listener (see, for example, Pfaff, 1979; McClure and Wentz, 1975).

Another distinction in switching from SL to PL and switching from PL to SL relates to the volitionality of the process. The latter type of code-switching, the sociolinguistic variety, is purposeful and voluntary. The former, our interest here, was by and large an involuntary, unintentional process resulting from SL semantic gaps or inability to access SL linguistic material. Indeed, even the phenomenon known in the literature as interference or interlanguage can be viewed as a form of code-switching within this analysis.

7.3 Selection of Measures and Stimulus Sampling

Two types of response scales can be identified in the relevant literature, one discrete (frequently dichotomous) and the other continuous. Measurement theories favor continuous scales. Among its advantages, a continuous scale can always be reduced to a discrete one, while the opposite is not necessarily the case. This way, whenever continuous measurement can be established, it is preferred. There have been quite a few studies in the field employing continuous measurement, e.g. reaction time.

The type of continuous measurement we used here, a rating scale, has demonstrated its effectiveness as the only unbiased scale for examining cognitive processes (see Anderson, 1982). Functional measurment takes an approach opposite to classical measurement theories such as Thurstone's methods. In functional measurement "scaling is derivative from substantive laws. As a consequence, primary concern shifts to response scaling" (Anderson, 1981:108). Anderson further states that due to "this orientation, stimulus scaling loses much of its explicit interest. In such designs, stimulus scaling may not be possible because the zeros and the units of the scale are arbitrary. Even three levels of a factor allow one nonarbitrary stimulus scale value."

Stimulus sampling is relevant here with regard to its mundane quality. To sample mundane content for the study of SL use, it would not be desirable to apply standard sampling procedures. Design of a second language experiment requires avoiding a temptation to rely on sampling procedures from PL studies (cf., Grosjean, 1985). Another requirement is to search for substantive cognitive units in the processing of SL (see Anderson, 1981:8-9). For an adequate examination of the integration prediction, stimulus sampling should relate to an inherent difference between two modes of language use, one unilingual (defining foreign lexical items in terms of the same language) and one bilingual (definitions in PL terms). We operationalized this in the use of two kinds of dictionaries to simulate SL processing.

Two lexical sources are available for constructing a dictionary for second language use. One is the PL; the other is the SL. There is a problem with regard to sampling of appropriate terms from dictionaries. In the priming study we sampled verbs. Verbs, however, may not facilitate the process we are in search of. In a sense, a verb as a hint item for an unfamiliar idiom captures only a portion of the semantic content of that idiom. Verbs such as 'leave' and 'escape' do not

express the prepositional element in an idiom like 'be off.' Similarly, verbs such as 'like' and 'tolerate' do not express the nominal element in an idiom like 'hit it off,' which can be glossed roughly as 'to become good friends' (Walters and Wolf, 1988).

In most of the experiments reported on above the actual stimulus level along a given dimension was a product of a similarity/ difference between a hint item and a target item. In the case of idiom comprehension, to simulate similarity we sampled hints which were lexical synonyms of the target item, while for differences we chose irrelevant words from the same grammatical category. In this way, the actual stimulus range was stretched to cover a maximal portion of the scale. While we gained with regard to scale range, we sacrificed internal validity, i.e. the task took on a somewhat artificial overtone. Future experimentation on these issues might explore a balance between these two requirements. Furthermore, it may be of value to conduct a series of experiments to test for possible roles of artificiality in studies of SL processing.

7.4 Future Perspectives

The interplay between first and second language use can be examined in a broader psychological context than the research reported above. Second language use can be described as a series of choices regarding lexical, syntactic and phonological options. The language user either has in mind a message to convey in his SL (production) or is challenged with a request to respond to a second language appeal. To fulfill such a task, from the moment he wishes to speak or respond, he presumably follows a set of rules and instructions which differ from his prevailing PL modus operandi about how to make lexical choices, formulate sentences, and produce utterances. In general terms, he makes a switch from a routine set of linguistic practices to another less assimilated and accomodated one. Such alternation is traditionally treated as a switch between two primary languages, e.g. English and Spanish.

There are other instances of switching from a commonly-used linguistic system to a secondary one. For example, a professional criminal has two different sub-languages generated from what second language researchers might call his mother tongue. The standard or formal language, as it is spoken, serves as a means of communication with people outside his criminal sub-culture. This language can be viewed as a second language, the first being the one used for communication within his criminal circle. According to this line of thought, in

research on bilingualism, a speaker who might be viewed as monolingual can be considered bilingual with regard to underlying psycholinguistic processing (see Paradis, 1980, and Gumperz, 1971, for a similar idea).

One can trace his own switching episodes between languages that differ with regard to formality. At one pole there is a primary language with a well-structured and well-formulated set of syntactic and phonological rules and lexical entries. To describe the other pole we might refer to Vygotsky's (1962) discussion of inner speech and its currrent revival in reading research (see Massaro, 1984). Inner speech makes use of idiosyncratic lexical networks (associated with the mother tongue of the language user). The nodes of these networks are assumed to be connected by idiosyncratic syntax. Inner speech has many gaps in formality as compared to the mother tongue it is linked with. These gaps, rather than leading to processing problems, are actually a source of greater communicative efficiency, since they are indicators of subjectively redundant elements, whether semantic or syntactic.

The compressed or reduced nature of inner speech is more efficient for fraternal or inner negotiation (i.e., thinking). When it comes to wider social negotiation, however, inner speech becomes inefficient in comparison to language based on more formal linguistic routines. For instance, a partially formal language which might be considered closest to inner speech is the talk of twins. A just noticeably wider social circle in terms of formality of language is the nuclear family. Next along this discrete range comes the criminal example presented above.

To what extent is a switch between formal English and criminal English like a switch between formal Spanish and formal English? For primary languages, like Spanish and English, formality and structure serve as guarantees of objectivity, i.e. they ensure equality of meaning for all those who share a particular linguistic code. An analysis of a switch between primary languages often indicates friction points in the phonology, syntax, and lexis. What we have called friction points revolves around notions such as word order and agreement and dependency relations. Nevertheless, primary languages frequently share notions of subject and object, or in structural terms, noun phrase and verb phrase.

A less formal language, e.g. criminal English, would not share this orderly structure with its primary or host language. Rather, it is elliptical, often leaving out major structural components. Taking this at face value, such a language does not have a logical syntax. It does not operate according to the definitional and functional basis of the formal logic that is fundamental to any primary language.

Sub-cultural or fraternal dependence encourages elimination of logical elements, rather than enforcement of formal syntax. Syntactically speaking, such a language can be conceived of as an assembling of structural codes. In order to decode them, one has to acquire the password for the momentarily agreed upon algorithm. The production rule for the algorithm is: 'Make it accessible to in-group members only.'

We can view a primary language as a formal and logical framework, which serves as a tool to enable a user to generate any sentence from among an infinite set of possibilities. Indeed, there are always common forms or routines that contemporary users share. The door is always open for consensus with regard to any particular set of forms as long as they follow the logical rules. To speak English or Spanish one does not have to know anybody personally, one does not have to join any fraternal organization, and one does not have to adopt any specific style or way of life, as for example criminals do. To communicate efficiently in a second language, a language user only needs to accomodate what he already knows about linguistic frameworks to the specifics of the new language.

A switch between primary and secondary languages is accompanied by a completely different sort of requirement which is derived from the special nature of informal languages. Here, rather than operating on the basis of a formal framework, one has to acquire a finite set of pragmatic terms. Such a switch can have no standard transfer rule, as is the case of a switch between two primary languages. In order to function appropriately in this linguistic context, it is necessary to get to know the speakers of that secondary language as well as the confidential pool of pragmatic symbols. It is a switch between a lexicon-dependent system and a syntactically-based framework.

This set of assumptions might lead to the prediction that primary (now SL)-to-secondary (now PL) code-switching is essentially more taxing than code-switching between two primary languages. This is because there are many more degrees of freedom needed to cross the former boundary. The analysis we have just proposed is unquestionably speculative, suffering on the one hand from the many untested assumptions and enjoying at the same time a high provocative coefficient.

It is easy to derive a series of working hypotheses from the above analysis, the most appealing aspect of which can be illustrated as follows: Consider a speaker of some primary language, e.g. English, being primed for switching in two directions, one from another primary language to English and the other from English to a secondary language. Such a comparison would reveal whether there

are identifiable points in secondary-to-primary switching, as is the case of primary-to-primary code-switching.

The presence of the 'right' person is what is usually required to trigger such a primary-to-secondary code-switch. This requirement goes with the whole routine involving identification of a person as belonging to the 'right' fraternal community. When it comes to primary-to-primary code-switching, on the other hand, most current research has examined linguistically-oriented switching between syntactic boundaries. Lexically speaking, a dictionary is an appropriate means to translate words from one primary language to another, but no dictionary can be constructed to capture the idiosyncratic, unstable nature of less formal languages.

REFERENCES

Aronson-Berman, R. (1975), "Analytic syntax: A technique for advanced level reading". *TESOL Quarterly 9*, 243-251.

Albert, M.K. and Obler, L.K. (1978), *The Bilingual Brain: Neuropsychological and Neurolinguistic Aspects of Bilingualism*. New York: Academic Press.

Alcalay, R. (1970), *The complete English-Hebrew Dictionary*. Tel Aviv: Massadah Publishing Company.

Andersen, R.W. (1983), "Transfer to somewhere". In: S. Gass and L. Selinker (Eds.), *Language Transfer and Language Learning*. Rowley, MA: Newbury House.

Andersen, R.W. (1984), "Discussant". In: A. Davies, C. Criper, and A.P.R. Howatt (Eds.), *Interlanguage*. Edinburgh: Edinburgh University Press.

Anderson, C.C. (1964), "The psychology of metaphor". *Journal of Genetic Psychology 105*, 53-73.

Anderson, N.H. (1981), *Foundations of Information Integration Theory*. New York: Academic Press.

Anderson, N.H. (1982), *Methods of Information Integration Theory*. New York: Academic Press.

Anderson, N.H. and Lopes, L.L. (1974), "Some psycholinguistic aspects of person perception". *Memory and Cognition 2*, 67-74.

Bailey, N., Madden, C., and Krashen, S. (1974), "Is there a natural sequence in adult second language learning". *Language Learning 21*, 235-243.

Barnhart, C.L. (1968), *The American College Dictionary*. New York: Random House.

Bates, E. and MacWhinney, B. (1979), "A functionalist approach to the acquisition of grammar". In: E. Ochs and B. Schieffelin (Eds.), *Developmental Pragmatics*. New York: Cambridge University Press.

Bates, E. and MacWhinney, B. (1981), "Second language acquisition from a functionalist perspective: Pragmatic, semantic, and perceptual strategies". In: H. Winitz (Ed.), *Native Language and Foreign Language Acquisition. Annals of the New York Academy of Sciences 379*, 190-214.

Bates, E. and MacWhinney, B. (1982), "Functionalist approaches to grammar". In: E. Wanner and L. Gleitman (Eds.), *Language Acquisition: The State of the Art*. New York: Cambridge University Press.

Bates, E. and MacWhinney, B. (1987), "Language universals, individual variation, and the competition model". In: B. MacWhinney (Ed.), *Mechanism of Language Acquisition*. Hillsdale, NJ: Erlbaum.

Bialystok, E. (1978), "A theoretical model of second language learning". *Language Learning 28*, 68-83.

Bialystok, E. (1981), "The role of linguistic knowledge in second language acquisition". *Studies in Second Language Acquisition 4*, 31-45.

Bialystok, E. (1984), "Cognitive dimensions of metalinguistic awareness". Paper presented at the annual meeting of the Canadian Psychological Association, Ottawa.

Bialystok, E. (1986), "Factors in the growth of linguistic awareness". *Child Development 57*, 498-510.

Billow, R.M. (1977), "Metaphor: A review of the literature". *Psychological Bulletin 84*, 81-92.

Black, M. (1962), "Metaphor". In: M. Black (Ed.), *Models and Metaphors*. Ithaca, NY: Cornell University Press.

Black, M. (1979), "More about metaphor". In: A. Ortony (Ed.), *Metaphor and Thought*. Cambridge: Cambridge University Press.

Brown, H.D. (1980), *Principles of Language Learning and Teaching*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Brown, H.D. (1987), *Principles of Language Learning and Teaching*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Brown, H.D., Crymes, R. and Yorio, C. (1977), *Teaching and Learning English as a Second Language: Trends in Research and Practice*. Washington, DC: TESOL.

Carrell, P. (1983a), "Three components of background knowledge in reading comprehension". *Language Learning 33*, 183-207.

Carrell, P. (1983b), "Background knowledge in second language comprehension". *Language Learning and Communication 2*, 25-33.

Carrell, P. (1983c), "Some issues in studying the role of schemata or background knowledge in second language comprehension". *Reading in a Foreign Language 1*, 81-92.

Carrell, P. (1984), "Evidence of a formal schema in second language comprehension". *Language Learning 34*, 87-112.

Carrell, P., Devine, J. and Eskey, D. (Eds.), (1988), *Interactive Approaches to Second Language Reading*. New York: Cambridge University Press.

Carroll, J.B. (1968), "Contrastive linguistics and interference theory". *19th Annual Georgetown University Roundtable*. Washington, D.C.: Georgetown University Press.

Chomsky, N. (1981), *Lectures on Government and Binding*. Dordrecht: Foris Publications.

Chomsky, N. (1986), *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger Publishers.

Clahsen, H. (1984), "The acquisition of German word order". In: R.W. Andersen (Ed.), *Second Languages: A Cross-Linguistic Perspective*. Rowley, MA: Newbury House.

Clahsen, H. (1987), "Connecting theories of language processing and (second) language acquisition". In: C.W. Pfaff (Ed.), *First and Second Language Processes*. Cambridge, MA: Newbury House.

Clark, H.H. and Clark, E.V. (1977), *Psychology and Language*. New York: Harcourt, Brace, and Jovanovich, Inc.

Cook, V.J. (1985), "Chomsky's universal grammar and second language learning". *Applied Linguistics 6*, 2-18.

Corder, S.P. (1967), "The significance of the learner's errors". *International Review of Applied Linguistics 5*, 161-170.

Corder, S.P. (1971), "Idiosyncratic dialects and error analysis". *International Review of Applied Linguistics 9*, 147-160.

Corder, S.P. (1975), "The language of second language learning: The broader issues". *Modern Language Journal 59*, 409-413.

Corder, S.P. (1973), *Introducing Applied Linguistics*. Baltimore, MD: Penguin Books Ltd.

Cowan, J.R. (1976), "Reading, perceptual strategies and contrastive analysis". *Language Learning 26*, 95-109.

Cziko, G. (1978), "Differences in first- and second-language reading: The use of syntactic, semantic and discourse constraints". *The Canadian Modern Language Review 34*, 473-489.

Dulay, H. and M. Burt (1974), "Natural sequences in child second language acquisition". *Language Learning 24*, 37-53.

Dulay, H., Burt, M. and Krashen, S. (1982), *Language Two*. New York: Oxford University Press.

Duskova, L. (1969), "On sources of error in foreign language learning". *International Review of Applied Linguistics 4*, 11-36.

Eckman, F. (1977), "Markedness and the constrastive analysis hypothesis". *Language Learning 27*, 315-330.

Eckman, F. (1985), "Some theoretical and pedagogical implications of the markedness differential hypothesis". *Studies in Second Language Acquisition 7*, 289-307.

Ervin, S. and Osgood, C.E. (1954), "Psycholinguistics: A survey of theory and research problems". *Journal of Abnormal and Social Psychology 49*, 139-146.

Faerch, K. and Kasper, G. (1986), "The role of comprehension in second-language learning". *Applied Linguistics 7*, 257-274.

Fillmore, L.W. (1979), "Individual differences in second language acquisition". In: C. Fillmore, D. Kempler, and W. Wang (Eds.), *Individual Differences in Language Ability and Language Behavior*. New York: Academic Press.

Flynn, S. (1984), "A universal in L2 acquisition based on a PBD typology". In: F. Eckman, L.H. Bell, and D. Nelson (Eds.), *Universals of Second Language Acquisition*. Rowley, MA: Newbury House.

Fodor, J. (1983), *Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Fraser, B. (1979), "The interpretation of novel metaphors". In: A. Ortony (Ed.), *Metaphor and Thought*. Cambridge: Cambridge University Press.

Fraser, B. (ms), *Second Language Acquisition: A Synthesis*. Boston University.

Fries, C. (1945), *Teaching and Learning English as a Foreign Language*. Ann Arbor: University of Michigan Press.

Gass, S. (1979), "Language transfer and universal grammatical relations". *Language Learning 29*, 327-344.

Gass, S. (1987), "The resolution of conflicts among competing systems: A bidirectional perspective". *Applied Psycholinguistics 8*, 329-350.

Gass and Ard, J. (1984), "The ontology of language universals". In: W.E. Rutherford (Ed.), *Language Universals and Second Language Acquisition*. Amsterdam: John Benjamins.

Greenberg, J.H. (Ed.), (1966), *Universals of Language*. Cambridge, MA: MIT Press.

Greuneich, R. (1982), "The development of children's integration rules for making moral judgements". *Child Development 53*, 887-894.

Grosjean, F. (1982), *Life with Two Languages*. Cambridge, MA: Harvard University Press.

Grosjean, F. (1985), "The bilingual as a competent but specific speaker-hearer". *Journal of Multilingual and Multicultural Development 6*, 467-477.

Grosjean, F. (ms). *Maman, tu peux me tailler mes chaussures? Exploring the recognition processes of guest words in bilingual* mixed speech. Northeastern University.

Gumperz, J.J. (1971), "On the linguistic markers of bilingual communication". *Journal of Social Issues 23*, 48-57.

Harris, R.J., Lahey, M.A., and Marsalek, F. (1980), "Metaphors and images: Rating, reporting, and remembering". In: R.P. Honeck and R.R. Hoffman (Eds.), *Cognition and Figurative Language*. Hillsdale, NJ: Erlbaum.

Honeck, R.P., Voegtle, K., Dorfmueller, M.A. and Hoffman, R. (1980), "Proverbs, meaning, and group structure". In: R.P. Honeck and R.R. Hoffman (Eds.), *Cognition and Figurative Language*. Hillsdale, NJ: Erlbaum.

Hyltenstam, K. (1982), "Markedness, language universals, language typology and second language acquisition". Paper presented at the Second European-North American Workshop on Cross-linguistic Second Language Acquisition Research, Gorhde, W. Germany.

Hyltenstam, (1983), "Data types and second language variability". In: H. Ringbom (Ed.), *Psycholinguistics and Foreign Language Learning*. Abo: Abo Akademi.

Ijaz, H. (1986), "Linguistic and cognitive determinants of lexical acquisition in a second language". *Language Learning 36*, 401-451.

Jackendoff, R. (1983), *Semantics and Cognition*. Cambridge, MA: MIT Press.

Jakobovitz, L. (1970), *Foreign Language Learning: A Psycholinguistic Analysis of the Issues*. Rowley, MA: Newbury House.

James, C. (1980), *Contrastive Analysis*. London: Longman.

Kail, M. (1987), "The development of sentence interpretation strategies from a cross-linguistic perspective". In: C.W. Pfaff (Ed.), *First and Second Language Processes*. Cambridge, MA: Newbury House.

Karmiloff-Smith, A. (1986), "Stage/structure versus phase/process in modelling linguistic and cognitive development". In: I. Levin (Ed.), *Stage and Structure: Reopening the Debate*. Norwood, NJ: Ablex.

Keller-Cohen, D. (1979), "Systematicity and variation in the non-native child's acquisition of conversational skills". *Language Learning 29*, 27-44.

Kellerman, E. (1977), "Towards a characterization of the strategy of transfer in second language learning". *Interlanguage Studies Bulletin 2*, 58-145.

Kellerman, E. (1978), "Giving learners a break: Native language intuition as a source of predictions about transferability". *Working Papers on Bilingualism 15*, 59-92.

Kellerman, E. (1979), "Transfer and non-transfer: Where we are now". *Studies in Second Language Acquisition 2*, 37-57.

Kellerman, E. (1983), "Now you see it, now you don't". In: S. Gass and L. Selinker (Eds.), *Language Transfer in Language Learning*. Rowley, MA: Newbury House.

Kellerman, E. (1984), "The empirical evidence for the influence of the L1 in interlanguage". In: A. Davies, C. Criper, and A.P.R. Howatt (Eds.), *Interlanguage*. Edinburgh: Edinburgh University Press.

Kilborn, K. and Cooreman, A. (1987), "Sentence interpretation strategies in adult Dutch-English bilinguals". *Applied Psycholinguistics 8*, 415-431.

Kintsch, W. (1974), *The Representation of Meaning in Memory*. Hillsdale, NJ: Erlbaum.

Kintsch, W. and van Dijk, T.A. (1978), "Toward a model of text comprehension and production". *Psychological Review 85*, 363-394.

Kintsch, W. (1988), "The role of knowledge in discourse comprehension: A construction-integration model". *Psychological Review 95*, 163-182.

Klein, W. (1986), *Second Language Acquisition*. London: Cambridge University Press.

Krashen, S. (1976), "Formal and informal linguistic environments in language learning and language acquisition". *TESOL Quarterly 10*, 157-168.

Krashen, S. (1977), "The monitor model of adult second language performance". In: M. Burt, H. Dulay, and M. Finocchiaro (Eds.), *Viewpoints on English as a Second Language*. New York: Regents.

Krashen, S. (1978), "Individual variation in the use of the monitor". In: W. Ritchie (Ed.), *Second Language Acquisition Research: Issues and Implications*. New York: Academic Press.

Krashen, S. (1981), *Second Language Acquisition and Second Language Learning*. Oxford: Pergamon Press.

Krashen, S. (1982), *Principles and Practice in Second Language Acquisition*. New York: Pergamon Press.

Krashen, S. and Pon, P. (1975), "An error analysis of an advanced ESL leaner. *Working Papers on Bilingualism 7*, 125-129.

Laberge, D. and Samuels, S.J. (1974), "Toward a theory of automatic processing in reading". *Cognitive Psychology 6*, 293-323.

Lado, R. (1957), *Linguistics Across Cultures: Applied Linguistics for Language Teachers*. Ann Arbor: University of Michigan Press.

Lado, R. (1968), "Contrastive analysis in a mentalistic theory of language learning". *19th Annual Georgetown University Roundtable*. Washington, D.C.: Georgetown University Press.

Lakoff, G. and Johnson, M. (1980), *Metaphors We Live By*. Chicago: University of Chicago Press.

Lambert, W.E. (1961), "Behavioral evidence for contrasting forms of bilingualism". *12th Annual Georgetown University Roundtable*. Washington, D.C.: Georgetown University Press.

Lamendella, J. (1977), "General principles of neurofunctional organization and their manifestation in primary and nonprimary language acquisition". *Language Learning 27*, 155-196.

Larsen-Freeman, D. (1975),"The acquisition of grammatical morphemes by adult ESL students". Paper presented at the *9th Annual TESOL Convention*, Los Angeles.

Larsen-Freeman, D. (1976), "An explanation for the morpheme acquisition order of second language learners". *Language Learning 26*, 125-134.

Leon, M. (1982), "Rules in children's moral judgements: Integration of intent, damage, and rational information". *Developmental Psychology 18*, 835-842.

Lesgold, A. and Perfetti, C.A. (1981), "Interaction processes in reading: Where do we stand?" In: A. Lesgold and C.A. Perfetti (Eds.), *Interactive Processes in Reading*. Hillsdale, NJ: Erlbaum.

LoCoco, V. (1975), "An analysis of Spanish and German learner'serrors". *Working Papers on Bilingualism 7*, 96-124.

Massaro, D. (1984), "Building and testing models of reading processes". In: D. Pearson (Ed.), *Handbook on Reading Research*. London: Longman.

MacWhinney, B. (1987a), "The competition model". In: B. MacWhinney (Ed.), *Mechanisms of Language Acquisition*. Hillsdale, NJ: Erlbaum.

MacWhinney, B. (1987b), "Applying the competition model to bilingualism". *Applied Psycholinguistics 8*, 315-327.

MacWhinney, B., Bates, E. and Kliegl, R. (1984), "Cue validity and sentence interpretation in English, German, and Italian". *Journal of Verbal Learning and Verbal Behavior 23*, 127-150.

McClure, E. (1977), "Aspects of code-switching in the discourse of bilingual Mexican-American children". *Georgetown University Roundtable on Languages and Linguistics*. Washington, D.C.: Georgetown University Press.

McClure, E. and Wentz, J. (1975), "Functions of code-switching among Mexican-Americans". In: *Papers from the Parasession on Functionalism*. Chicago: Chicago Linguistic Society.

McDonald, J.L. (1986), "The development of sentence comprehension strategies in English and Dutch". *Journal of Experimental Child Psychology 41*, 317-335.

McDonald, J.L. (1987), "Sentence interpretation in bilingual speakers of English and Dutch". *Applied Psycholinguistics 8*, 379-413.

McKormack, P.D. (1977), "Bilingual linguistic memory: The independence-interdependence issue revisited". In: P.A. Hornby (Ed.), *Bilingualism: Psychological, Social, and Educational Implications*. New York: Academic Press.

McLaughlin, B. (1987), *Theories of Second Language Learning*. Cambridge: Edward Arnold.

McLaughlin, B., Rossman, T. and McLeod, B. (1983), "Second language learning: An information processing perspective". *Language Learning 33*, 135-158.

Meisel, J. (1980), "Strategies of second language acquisition". In: R.W. Andersen (Ed.), *Pidginization and Creolization as Language Acquisition*. Rowley, MA: Newbury House.

Miller, G.A. (1979), "Images and models, similes and metaphors". In: A. Ortony (Ed.), *Metaphor and Thought*. Cambridge: Cambridge University Press.

Oden, G. (1977), "Fuzziness in semantic memory". *Memory and Cognition 5*, 198-204.

Oden, G. (1978), "Semantic constraints and judged preference for interpretations of ambiguous sentences". *Memory and Cognition 6*, 26-37.

Oden, G. (1986), "On the use of semantic constraints in Guiding syntactic analysis". *International Journal of Man-Machine Studies 19*, 335-357.

Oden, G. and Anderson, N.H. (1974), "Integration of semantic constraints". *Journal of Verbal Learning and Verbal Behavior 13*, 138-148.

Oller, J.W. and Richards, J.C. (Eds.), (1973), *Focus on the Learner*. Rowley, MA: Newbury House.

Ortony, A. (1979), "The role of similarity in similes and metaphors". In: A. Ortony (Ed.), *Metaphor and Thought*. Cambridge: Cambridge University Press.

Ortony, A. (1980), "Some psycholinguistic aspects of metaphor". In: R.P. Honeck and R.R. Hoffman (Eds.), *Cognition and Figurative Language*. Hillsdale, NJ: Erlbaum.

Paivio, A. (1979), "Psychological processes in the comprehension of metaphor". In: A. Ortony (Ed.), *Metaphor and Thought*. Cambridge: Cambridge University Press.

Paradis, M. (1980), "The language switch in bilinguals". In: P. Nelde (Ed.), *Languages in Contact and in Conflict*. Wiesbaden: Franz Steiner Verlag.

Perrine, R. (1971), Four forms of metaphor". *College English 33*, 125-138.

Pfaff, C. (1979), "Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish-English". *Language 55*, 291-318.

Pfaff, C. (1984), "On input and residual L1 transfer effects in Turkish and Greek children's German". In: R.W. Andersen (Ed.), *Second Languages: A Cross-linguistic Perspective*. Rowley, MA: Newbury House.

Phinney, M. (1987), "The pro-drop parameter in second language acquisition". In: T. Roeper and E. Williams (Eds.), *Parameter Setting*. Dordrecht: D. Reidel Publishing Company.

Poplack, S. and Sankoff, D. (1987), "Code-switching". In: U. Ammon, N. Dittmar, and Klauss J. Mattheier (Eds.), *Sociolinguistics: An International Handbook of the Science of Language and Society*. Berlin: Walter de Gruyter.

Reichmann, P.P. and Coste, E.L. (1980), "Mental imagery and the comprehension of figurative language. Is there a relationship?" In: R.P. Honeck and R.R. Hoffman (Eds.), *Cognition and Figurative Language*. Hillsdale, NJ: Erlbaum.

Richards, I.A. (1936), *The Philosophy of Rhetoric*. London: Oxford University Press.

Ritchie, W.C. (1967), "Some implications of generative grammar for the construction of courses in English as a foreign language". *Language Learning 17*, 45-69.

Rosansky, E. (1976), "Methods and morphemes in second language acquisition". *Language Learning 26*, 409-425.

Rumelhart, D. (1977), "Toward an interactive model of reading". In: S. Dornic (Ed.), *Attention and Performance*, VI. Hillsdale, NJ: Erlbaum.

Rumelhart, D.E. (1979), "Some problems with the notion of literal meanings". In: A. Ortony (Ed.), *Metaphor and Thought*. Cambridge: Cambridge University Press.

Rumelhart, D.E., McClelland, J.L., and the PDP Research Group. (1986), *Parallel Distributive Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press.

Rutherford, W.E. (1985), "Grammatical theory and L2 acquisition: A brief overview". *Second Language Research 2*, 1-11.

Rutherford, W.E. (ms). *Questions of learnability in second language acquisition*. University of Southern California.

Schachter, J. (1974), "An error in error analysis". *Language Learning 24*, 205-214.

Schumann, J. (1981), "Discussion of two perspectives on pidginisation as second language acquisition". In: R.W. Andersen (Ed.), *Pidginization and Creolization as Language Acquisition*. Rowley, MA: Newbury House.

Searle, J. (1979), "Metaphor". In: A. Ortony (Ed.), *Metaphor and Thought*. Cambridge: Cambridge University Press.

Segalowitz, N. (1977), "Psychological perspectives on bilingual education". In: B. Spolsky and R. Cooper (Eds.), *Frontiers in Bilingual Education*. Rowley, MA: Newbury House.

Sharwood Smith, M. (1986), "Comprehension versus acquisition: Two ways of processing input". *Applied Linguistics 7*, 239-256.

Shiffrin, R.M. and Schneider, W. (1977), "Controlled and automatic human information processing. II: perceptual learning, automatic attending and a general theory". *Psychological Review 84*, 127-190.

Spolsky, B. (1986), "Formulating a theory of second language learning". *Studies in Second Language Acquisition 7*, 269-288.

Stein, N.L. and Glenn, C.G. (1979), "An analysis of story grammar comprehension in elementary school children". In: R.O. Freedle (Ed.), *New Directions in Discourse Processing, Volume 2*. Norwood, NJ: Ablex.

Tarone, E., Cohen, A. and Dumas, G. (1976), "A closer look at some interlanguage terminology". *Working Papers on Bilingualism 9*, 76-90.

Ulijn, J. and Kempen, G.A.M. (1976), "The role of the first language in second language reading comprehension: Some experimental evidence". In: G. Nickel (Ed.), *Proceedings of the 4th International Congress of Applied Linguistics*. Stuttgart: Hochschul Verlag.

Ulijn, J. (1979), "L'apprentissage de la lecture du français scientifique et la langue maternelle". *Working Papers on Bilingualism 19*, 89-114.

Ulijn, J. (1980), "Foreign language reading research: Recent trends and future prospects". *Journal of Research in Reading 3*, 17-37.

Ulijn, J. (1981), "Conceptual and syntactic strategies in reading a foreign language". In: E. Hopkins and R. Grotjahn (Eds.), *Studies in Language Teaching and Language Acquisition*. Bochum: Brockmeyer.

Vogt, H. (1949), "Dans quelles conditions et dans quelles limites peut s'exercer sur le systeme morphologique d'une autre langue?" *International Congress of Linguists 231*, 31-45.

Vygotsky, L. (1962), *Thought and Language*. Cambridge, MA: MIT Press.

Walters, J. and Y. Wolf (1988), "Integration of first language material in second language comprehension". In: J. Fine (Ed.), *Second Language Discourse*. Norwood, NJ: Ablex.

Wardaugh, R. (1970), "The contrastive analysis hypothesis". *TESOL Quarterly 4*, 123-130.

Weinreich, U. (1967), *Languages in Contact*. The Hague: Mouton. (Originally published by the New York Linguistic Circle, 1953.)

White, L. (1983), "Markedness and parameter setting: Some implications for a theory of adult second language acquisition". In: F. Eckman, E. Moravcsik, and J. Wirth (Eds.), *Proceedings of the 12th Annual University of Wisconsin-Milwaukee Symposium on Markedness*. New York: Plenum Press.

White, L. (1984), "Implications of parametric variation for adult second language acquisition: An investigation of the pro-drop parameter". In: V. Cook (Ed.), *Experimental Approaches to Second Language Acquisition.* Oxford: Pergamon Press.

White, L. (1985), "The pro-drop parameter in adult second language acquisition". *Language Learning 35*, 47-62.

Winner, E. (1979), "New names for old things". *Journal of Child Language 6*, 469-491.

Wode, H. (1981), *Learning a Second Language*. Tübingen: Gunter Narr Verlag.

Wolf, Y. and Walters, J. (1988), "Second language comprehension processes: Theories and methods". In: J. Fine (Ed.), *Second Language Discourse.* Norwood, NJ: Ablex.

Wolf, Y., Walters, J. and Holzman, S. (1989), "Integration of semantic and structural constraints in narrative comprehension". *Discourse Processes 12*, 1449-167.

Wolfe, D.L. (1967), "Some theoretical aspects of language learning and language teaching". *Language Learning 17*, 173-188.

Yorio, C.A. (1971), "Some sources of reading problems for foreign language learners". *Language Learning 21*, 107-115.

Zobl, H. (1982), "A direction for constrastive analysis: The comparative study of developmental sequences". *TESOL Quarterly 16*, 169-183.

Zobl, H. (1983), "Markedness and the projection problem". *Language Learning 33*, 293-313.

Zobl, H. (1984), "Cross-language generalizations and the contrastive dimension of the interlanguage hypotheses". In: A. Davies, C. Criper, and A.P.R. Howatt (Eds.), *Interlanguage*. Edinburgh: Edinburgh University Press.

Rieger, Burghard (ed.)
Glottometrika 13
Bochum: Brockmeyer 1992, 105-120

Two models for word association data[1]

Gabriel Altmann

1. Though the treatment of word association data has changed somewhat in recent years (cf. Matthäus 1980), the fact that they can be ascertained empirically remains unchanged. From the linguistic point of view it is interesting to examine them on two grounds:

(i) A given word may have various associations. Some of them represent general connotations of the word, accepted by the given language community; they are the source of word dynamics (the creation of synonymy, homonymy, polysemy etc.) and the object of synchronic and historical semantics. They are distinguished by a high frequency of occurrence. Others, characterized by low frequencies, consist of idiolectal, "private" associations, the examination of which can be relevant in psychiatry.

(ii) Word associations represent a particular dimension of the diversifications of a word (cf. Altmann 1989a). Their ranked frequencies of occurrence show a regular pattern stimulating us to examine its form and causes. Their investigation is relevant for "synergetic linguistics" concerned with language processes and language self-regulation.

More than 25 years ago Horvath (1963) examined this pattern for the first time and concluded that it followed the Yule distribution. Though its derivation by Simon (1955) did not refer to the ranking of word associations, the reasoning by Haight and Jones (1974) (cf. also Lánsky, Radil-Weiss 1980) seemed to be plausible. Unfortunately, our fittings of this distribution to the data of Palermo

[1] This study is part of the project "Language Synergetics" sponsored by the Volkswagen Foundation.

and Jenkins (1964) showed that only about 50% of the empirical distributions tested followed it. Haight (1966) derived for fitting the ranked data the zeta-distribution but the cases tested by us brought still worse results than the Yule distribution. The same holds for the Borel and the logarithmic series distributions tested by Haight (1966). For illustration see the fitting of these distributions in Table 1.

Table 1. Fitting some distributions to the word-associations of "high", 4th grade, male (Palermo, Jenkins)

x	f_x	Logseries	Yule	Borel	Haight-Zeta
1	129	134.71	102.03	100.81	112.61
2	16	47.21	36.64	36.51	31.73
3	14	22.06	19.36	19.84	16.14
4	12	11.60	12.14	12.77	10.08
5	6	6.50	8.39	9.03	7.02
6	5	3.80	6.18	6.78	5.23
7	4	2.28	4.77	5.31	4.08
8	4	1.40	3.80	4.29	3.29
9	3	0.87	3.10	3.54	2.72
10	3	0.55	2.59	2.98	2.30
11	3	0.35	2.20	2.54	1.97
12	2	0.23	1.89	2.20	1.71
13	2	0.15	1.65	1.92	1.51
14	2	0.09	1.45	1.69	1.34
15	2	0.06	1.28	1.50	1.20
16	2	0.04	1.15	1.34	1.08
17	2	0.03	1.03	1.21	0.98
18	2	0.02	0.93	1.09	0.89
19	2	0.01	0.85	0.99	0.82
20	2	0.01	0.78	0.90	0.75
21	1	0.01	0.71	0.83	0.70
22	1	0.00	0.66	0.76	0.65

x	f_x	Logseries	Yule	Borel	Haight-Zeta
23	1	0.00	0.61	0.70	0.60
24	1	0.00	0.56	0.65	0.56
25	1	0.00	0.53	0.60	0.53
26	1	0.00	0.49	0.56	0.49
27	1	0.00	0.46	0.52	0.47
28	1	0.00	0.43	0.48	0.44
29	1	0.00	0.40	0.45	0.42
30	1	0.00	0.38	0.42	0.39
31	1	0.00	0.36	0.40	0.37
32	1	0.00	0.34	0.37	0.35
33	1	0.00	0.32	0.35	0.34
34	1	0.00	0.31	0.33	0.32
35	1	0.00	13.23	7.33	17.93
Parameter:		0.7010	0.7851	0.8335	0.6047
χ^2:		676.94	50.29	40.39	45.26
df:		7	22	23	22
P:		$<10^{-10}$	0.0005	0.0139	0.0025

Borel d. $$P_x = \frac{e^{-ax}a^{x-1}x^x}{(x-1)!}, \qquad x = 1,2,3,\ldots$$

Logseries d. $$P_x = \frac{q^x}{-x\ln(1-q)}, \qquad x = 1,2,3,\ldots$$

Yule d. $$P_x = \frac{bx!}{(b+1)^{(x+1)}}, \qquad x = 0,1,2,\ldots\text{(shifted)}$$

Haight-Zeta d. $$P_x = \frac{1}{(2x-1)^a} - \frac{1}{(2x+1)^a}, \qquad x = 1,2,3,\ldots$$

These and the following fittings were performed with the program FITTER that begins with the classical estimators (maximum likelihood, method of mo-

ments, method of frequency classes, least squares etc.) and improves the fitting iteratively in order to obtain the minimum chi-square value. The program contains about 200 different discrete probability distributions.

2. The models developed so far sought a purely probabilistic mechanism generating the data and left the specificity of language aside. Though chance is omnipresent in language, it is not the only cause of change and not the only background for existing structures. In the last decade a trend to use in linguistics some ideas of G.K.Zipf has developed. One possibility is to generalize the Zipf distribution (Zipf-Estoup d., discrete Pareto d., Riemann zeta d.). This is the way chosen by V.A. Dolinskij (1988), who used the function

$$f_x = f_1 x^{-(a+b \ln x)} \qquad x = 1, 2, 3, \ldots \tag{1}$$

In order to make of (1) a probability distribution we redefine it as follows:

$$P_x = \begin{cases} \alpha & x = 1 \\ \dfrac{\alpha x^{-(a+b \ln x)}}{T} & x = 2, 3, \ldots, n \end{cases} \tag{2}$$

with $\alpha^* = f_1 / N$ and $T = \alpha \sum_{i=2}^{n} P_i / (1-\alpha)$

Table 2. Fitting the Zipf-Dolinskij distribution to the word association of "high", 4th grade, male

x	f_x	NP_x
1	129	129.00
2	16	18.29
3	14	11.89
4	12	8.73
5	6	6.86
6	5	5.62
7	4	4.75
8	4	4.10
9	3	3.60

x	f_x	NP_x
10	3	3.20
11	3	2.88
12	2	2.61
13	2	2.39
14	2	2.20
15	2	2.04
16	2	1.89
17	2	1.77
18	2	1.66
19	2	1.56
20	2	1.47
21	1	1.39
22	1	1.32
23	1	1.26
24	1	1.20
25	1	1.14
26	1	1.09
27	1	1.05
28	1	1.00
29	1	0.96
30	1	0.93
31	1	0.89
32	1	0.86
33	1	0.83
34	1	0.80
35	1	0.70
$a = 1.0323$	$b = 0.0169$	
$\chi^2 = 3.43$;	$df = 27$; $P = 1.0$	

The result of fitting (2) to the same data as above is presented explicitly in Table 2. As can be seen, the fitting is excellent and the same holds for all other data in Table 3, where only the parameters and the test results are presented. For the sake of comparison we chose from Palermo and Jenkins the same words as Haight,

namely "music", "blossom", "table" and "high". In the majority of cases the probability of the resulting chi-square was 1.00 and there was no significant deviation from this model. It should be remarked that most cases of Table 3 can be successfully fitted with one positive and one negative parameter (a,b). We have chosen the best results yielded by the programm FITTER. Thus the parameters in Table 3 are not suitable for further (linguistic or psychological) interpretation.

Though the results are excellent, the present author could not find any linguistic basis for (2) so that there still remains the problem of systematization which is a necessary condition of theory building.

Table 3. Fitting the Zipf-Dolinskij distribution to the associations of "table", "music", "blossom", and "high" from Palermo, Jenkins

"table"								
Grade	4		5		6		7	
Gender	M	F	M	F	M	F	M	F
a	1.47151	.5073	.9264	.91141	.22631	.00171	.5137	1.3121
b	.0002	.0001	.0867	.1791	.0293	.1301	.0001	.00005
χ^2	3.32	6.64	8.26	10.80	2.79	4.86	5.78	2.69
df	19	20	24	18	24	19	20	18
P	1.00	1.00	1.00	.90	1.00	1.00	1.00	1.00,

"table"								
Grade	8		10		12		College	
Gender	M	F	M	F	M	F	M	F
a	1.8470	2.5365	.3376	.7366	.7538	.9951	.6005	.7099
b	.00005	-.2623	.2875	.1494	.1491	.1430	.2017	.1404
χ^2	14.45	1.28	7.58	5.94	3.07	1.41	7.08	.91
df	14	16	19	18	16	13	24	21
P	.42	1.00	.99	1.00	1.00	1.00	1.00	1.00

"high"								
Grade	4		5		6		7	
Gender	M	F	M	F	M	F	M	F
a	1.0323	1.2067	.4013	.7812	.9761	.85261	.06831	.1471
b	.0169	.0001	.2094	.1325	.0844	.0905	.0311	.0779
χ^2	3.43	4.32	5.80	5.78	8.70	4.09	3.96	3.50
df	27	27	25	23	24	25	28	20,
P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00,

"high"								
Grade	8		10		12		College	
Gender	M	F	M	F	M	F	M	F
a	.4668	1.2020	1.0518	1.1962	1.2128	.5678	.4246	1.2257
b	.1330	.0013	.0588	.0003	.0002	.1227	.1702	.0116
χ^2	1.43	3.31	5.85	2.30	3.49	2.13	3.55	2.85
df	25	22	27	27	25	26	35	34
P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

"music"								
Grade	4		5		6		7	
Gender	M	F	M	F	M	F	M	F
a	1.1484	.1584	.1655	.1462	.15721	.0011	.7369	.4665
b	.0003	.2091	.2402	.2564	.2150	.0453	.0877	.1469
χ^2	3.95	3.93	4.89	7.68	4.46	4.13	3.94	15.06
df	40	35	35	35	37	40	39	39
P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

"music"								
Grade	8		10		12		College	
Gender	M	F	M	F	M	F	M	F
a	.3131	.3860	.0798	.0407	.0547	.8406	1.2500	1.3501
b	.1630	.1850	.1932	.2626	.1735	.0764	.0001	.0001
χ^2	7.34	4.95	3.96	3.92	2.57	5.74	7.92	34.89
df	42	36	43	34	48	37	55	53
P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.97,

"blossom"								
Grade	4		5		6		7	
Gender	M	F	M	F	M	F	M	F
a	.0046	.8904	.7204	.33522	.74971	.0784	2.1413	.4289
b	.2050	.0503	.0454	.2410	-.4770	.0855	-.2803	.2608
χ^2	1.19	4.57	2.50	.95	9.67	2.51	1.90	2.76
df	15	13	15	12	20	9	15	13
P	1.00	.98	1.00	1.00	.97	.98	1.00	1.00

"table"								
Grade	8		10		12		College	
Gender	M	F	M	F	M	F	M	F
a	.0004	.6359	1.2019	1.4652	2.0577	1.3590	.9610	1.5134
b	.1798	.2317	.0002	.0013	-.2359	.0008	.0427	.00004
χ^2	2.52	1.21	2.51	3.09	2.49	2.37	2.72	3.08
df	16	11	17	14	23	15	29	23
P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

3. For this reason let us use another idea of Zipf's, namely that of the "Zipfian forces" of diversification and unification (cf. Zipf 1949), which together with Köhler's communicative needs of language users (cf. Köhler 1986) have found their place in "synergetic linguistics". The Zipfian forces can be interpreted as a consequence of the strivings of the speaker and the hearer for least effort in communication, while Köhler's "needs" represent a more general class of linguistic factors.

Starting from Zipf's dynamics we can say that e.g. in the domain of semantics the hearer strives to eliminate polysemy and allow each word to have exactly one denotation. This measure enables him (among other things) to receive the message with the least decoding effort. The reverse case, namely that each word has as many meanings as possible, causes the least coding effort for the speaker. In the extreme case this "victory" of the speaker would lead to the elimination of all but one single word, which would convey all meanings. The empirical result of these contrary tendencies is a compromise leading to a special probability distribution of meanings.

Various "Zipfian processes" of this kind shape the form and the meaning of the word, one of them being the creation of word associations. We are not interested in the problem of how and when an association arises. Our problem is the form of their probability distribution and its linguistic foundation.

The tendency of the hearer in respect to word associations is the same as with denotations: he tries to minimize their number. This renders his understanding of the message easier than in the case where he is forced to reflect the "hidden" meaning of each word. The speaker, on the contrary, in order to satisfy his needs for expression (cf. Bühler 1965; Köhler 1986), incessantly generates new word associations, new con-notations. They can be different with every speaker, and their number and frequency can be different with every word. This means that the semantics of the word undergoes incessant diversification and unification. The result is necessarily a compromise generating a skew distribution in which one or few associations are very frequent and the tail is usually very long. Since ranking is a special ordering or even a special view of data (cf. Arapov, Efimova, Srejder 1975; Arapov 1988; Altmann 1989b) it is not necessary for a model describing this course to be essentially monotonic decreasing. It is sufficient that for special values of parameters it may take this form. As a matter of fact, the probability distribution of word associations can be conceived in another way if one groups the associations according to another measurable criterion than rank.

Let us assume that word associations have always existed and that their distribution has always followed the same model in spite of numerous reshufflings, word and test person differences. This assumption is plausible because we take the existence of a unique, balancing self-regulation law for granted. It means then that there is an equilibrium distribution that can be ascertained from the relations of the neighbouring frequency classes.

Following this direction we can postulate that the probability of the rank x, P_x, is proportional to P_{x-1}, i.e. $P_x \sim P_{x-1}$, or that the difference $P_x - P_{x-1} \sim P_{x-1}$. This proportionality should contain both the di-versification tendency of the speaker and the unification tendency of the hearer. Thus regarding the second alternative we simply assume that the hearer exerts the unification pressure ax from which the constant influence b of the speech community must be subtracted while the speaker diversifies the association field with the force cx. The difference $P_x - P_{x-1}$ is of course negative with ranking. Putting all these factors together we obtain

$$P_x - P_{x-1} = -\frac{ax-b}{cx} P_{x-1} \quad , \qquad x \geq 1 \qquad (3)$$

Table 4. Fitting the negative binomial distribution to the word association of "high", 4th grade, male

x	f_x	NP_x
1	129	126.50,
2	16	23.62
3	14	13.48
4	12	9.42
5	6	7.19
6	5	5.76
7	4	4.76
8	4	4.03
9	3	3.46
10	3	3.01
11	3	2.64
12	2	2.34
13	2	2.08
14	2	1.87
15	2	1.68
16	2	1.52
17	2	1.38
18	2	1.25
19	2	1.14
20	2	1.05
21	1	0.96
22	1	0.88
23	1	0.81
24	1	0.75
25	1	0.69
26	1	0.64
27	1	0.59

x	f_x	NP_x
28	1	0.55,
29	1	0.51
30	1	0.47,
31	1	0.44
32	1	0.41
33	1	0.38
34	1	0.35
35	1	5.39
$k = 0.1955$;	$p = 0.0499$	
$\chi^2 = 15.78$;	$df = 24$; $P = 0.90$	

The greater the denominator the smaller the difference, yielding a long tail. It can easily be seen that (3) is a special case of the system of Ord (1967, 1972). Writing

$(c-a)/c = q$ and $b/(c-a) = k - 1$

we obtain as the solution of (3) the negative binomial distribution

$$P_x = \binom{k+x-1}{x} p^k q^x \qquad x = 0, 1, \ldots \qquad (4)$$

Since ranking can begin with 0 or 1 we can shift (4) one step to the right, in order to produce agreement with the usual ranking convention.

The fitting of the negative binomial distribution to the data is shown in Tables 4 and 5. In Table 4 the fitting to the associations of "high" (4th grade, male) is presented explicitly, in order to enable the reader to compare it with the results in Table 1. In Table 5 we present only the parameters and the results of testing.

It can be seen that in Table 5 only one case among 48 ("music", College, female) shows a significantly high deviation from the empirical data; two cases ("table", 8th grade, male; "blossom", 8th grade, male) have $P = .01$ and $P = .02$ respectively; all the others show an excellent agreement.

As a matter of fact these results are not as good as those using the Zipf-Dolinskij model. Nevertheless they are satisfying. So we have the possibility of choosing one of the two models for special purposes: For practical (descriptive) purposes and even for the examination of dependencies of associations on age or sex of speakers or on some properties of the words (length, frequency, age etc.) the Dolinskij model can render good service until its embedding within a theory becomes possible.

Table 5. Fitting the negative binomial distribution to the associations of "table", "music", "blossom", and "high" from Palermo, Jenkins

"table"								
Grade	4		5		6		7	
Gender	M	F	M	F	M	F	M	F
k	.2898	.3109	.3549	.3426	.3763	.3393	.2721	.1854
p	.0722	.0775	.0702	.0889	.0706	.1026	.0885	.0624
χ^2	8.82	16.39	18.18	16.31	8.75	8.87	12.03	8.43
df	20	21	23	20	23	18	19	18
P	.98	.75	.75	.70	.99	.96	.88	.97

"table"								
Grade	8		10		12		College	
Gender	M	F	M	F	M	F	M	F
k	.3019	.1866	.2921	.1640	.1802	.1372	.1011	.1476
p	.1002	.0767	.0865	.0641	.0720	.0840	.0467	.0622,
χ^2	33.16	9.86	7.63	11.39	9.62	5.30	13.70	16.12
df	17	17	11	18	16	13	20	23
P	.01	.91	.99	.88	.89	.97	.85	.85

"music"								
Grade	4		5		6		7	
Gender	M	F	M	F	M	F	M	F
k	.5805	.6865	.6423	.5970	.6551	.5620	.5643	.5251
p	.0523	.0705	.0668	.0700	.0601	.0592	.0570	.0432
χ^2	23.87	11.36	15.03	20.51	14.72	22.95	16.83	30.33
df	37	33	34	34	36	36	36	40
P	.95	1.00	1.00	.97	1.00	0.96	1.00	.87

"music"								
Grade	8		10		12		College	
Gender	M	F	M	F	M	F	M	F
k	.5391	.1141	.6060	.6413	.5534	.4966	.4930	.5104
p	.0410	.0346	.0458	.0648	.0425	.0572	.0510	.0542
χ^2	19.20	24.53	12.83	12.15	18.98	18.21	41.15	107.53
df	40	33	41	34	43	35	47	47
P	1.00	.86	1.00	1.00	1.00	.99	.70	$12u10^{-7}$

"blossom"								
Grade	4		5		6		7	
Gender	M	F	M	F	M	F	M	F
k	.1083	.1253	.1285	.1221	.1041	.1147	.1511	.1409
p	.0586	.0870	.0680	.0692	.0432	.1118	.0727	.0814
χ^2	21.47	13.87	17.92	10.61	16.12	7.37	8.32	6.66
df	13	11	13	12	15	9	13	13
P	.06	.24	.16	.56	.37	.60	.82	.92

"blossom"								
Grade	8		10		12		College	
Gender	M	F	M	F	M	F	M	F
k	.1057	.1302	.1509	.1929	.1524	.1479	.1307	.1686
p	.0358	.0993	.0666	.1027	.0550	.0792	.0444	.0761
χ^2	27.30	5.26	11.45	6.16	10.99	6.87	21.36	9.22
df	14	11	16	13	18	14	25	21
P	.02	.92	.78	.94	.89	.94	.67	.99

"high"								
Grade	4		5		6		7	
Gender	M	F	M	F	M	F	M	F
k	.1955	.2016	.2583	.2254	.3034	.2669	.2568	.2741
p	.0449	.0454	.0574	.0603	.0700	.0596	.0528	.0844
χ^2	15.78	13.93	10.78	12.13	15.65	12.04	13.02	8.18
df	24	25	24	22	23	24	25	20
P	.90	.96	.99	.95	.87	.98	.98	.99

"high"								
Grade	8		10		12		College	
Gender	M	F	M	F	M	F	M	F
k	.1917	.1730	.2653	.2301	.2340	.1983	.1713	.1538
p	.0475	.0529	.0547	.0508	.0581	.0465	.0398	.0428
χ^2	17.97	12.01	14.04	11.94	12.91	14.97	22.98	13.55
df	22	20	25	24	22	23	32	30
P	.71	.92	.96	.98	.94	.90	.88	.99

The other model is more suitable for explanation since it fits the explanandum into a discernible nomological pattern (s. Salmon 1984:17 f.), shows the factors leading to the given form, reveals a local self-regulation and puts this phenomenon into the broad family of diversification phenomena in language.

Conclusion

In models of linguistic probability distributions at least the local self-regulation - in form of a steady-state distribution - should be taken into account. In our case it is the competition between the needs of speakers and hearers which leads to the given distribution and yields both a good description and an acceptable explanation.

Since language phenomena are never isolated, it would be necessary to examine (i) how the parameters of the individual distributions correlate with other properties of the stimulus words or (ii) with the properties of the speakers, (iii) whether they are perhaps statistically equal and (iv) whether grouping of the response words according to other criteria would yield other distributions.

References

Altmann, G. (1989), "Diversification processes of the word". In: Köhler, R. (ed.), *Studies in Language Synergetics* (to appear).

Altmann, G. (1991), "Modelling diversification phenomena in language". In: Rothe, U. (ed.), *Diversification processes in language: Grammar* . Hagen, 33-46.

Arapov, M.V. (1988), *Kvantitativnaja lingvistika*. Moskva: Nauka.

Arapov, M.V., Efimova, E.N., Srejder, Ju.A. (1975), "O smysle rangovych raspredelenij". *Naucno-techniceskaja informacija ser.* 2, Nr. 1, 9-20.

Bühler, K. (1965), *Sprachtheorie*. Stuttgart: Fischer.

Dolinskij, V. A. (1988), "Raspredelenie reakcij v experimentach po verbal'nym associacijam". *Acta et Commentationes Universitatis Tartuensis* 827, 89-101.

Haight, F.A. (1966), "Some statistical problems in connection with word association data". *Journal of Mathematical Psychology* 3, 217-233.

Haight, F.A., Jones, R.B. (1974), "A probabilistic treatment of quantitative data with special reference to word association tests". *Journal of Mathematical Psychology* 11, 237-244.

Horvath, W.J. (1963) "A stochastic model for word association tests". *Psychological Review* 70, 361-364.

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

Lánsky, P., Radil-Weiss, T. (1980), "A generalization of the Yule-Simon model, with special reference to word association tests and neural cell assembly formation". *Journal of Mathematical Psychology* 21, 53-65.

Matthäus, W. (1980), *Theoretische und experimentelle Untersuchungen zum verbalen Assoziieren. Methodenkritisch gerichtete Strukturanalysen von Einfallsreihen und ihrer Veränderung bei Aufgabenwechsel*. Bochum: Brockmeyer.

Ord, J.K. (1967), "On a system of discrete distributions". *Biometrika* 54, 649-656.

Ord, J.K. (1972), *Families of frequency distributions*. London: Griffin.

Palermo, D.S., Jenkins, J.J. (1964), *Word Association Norms. Grade School through College*. Minneapolis: University of Minnesota Press.

Salmon, W.C. (1984), *Scientific explanation and the causal structure of the world*. Princeton, New Jersey: Princeton University Press.

Simon, H.A. (1955), "On a class of skew distribution functions". *Biometrika* 42, 425-440.

Zipf, G.K. (1949), *Human behavior and the principle of least effort*. Reading, Mass.: Addison-Wesley.

Rieger, Burghard (ed.)
Glottometrika 13
Bochum: Brockmeyer 1992, 121–172

Evaluating the adequacy of regression models: Some potential pitfalls*

Rüdiger Grotjahn

1. Introduction

It is characteristic of quantitative and especially of synergetic linguistics that quantitative models are used for both theory building and the description, prediction and explanation of data. More specifically, the following purposes may be involved in the use of quantitative models (cf. Daniel, Wood 1980: 5; Rayner, Best 1989: 4):

- summarizing a mass of data, for example for the purpose of prediction or control
- confirming or rejecting a theoretical relationship
- shedding light on the mechanisms generating the data
- comparing several sets of data in terms of the constants of the fitted models
- helping to choose a theoretical model
- providing a model which can be used in parametric inferential statistics.

Quantitative models may be derived in two ways. In the first approach, which is methodologically preferable and which is prevalent in synergetic linguistics, one tries to derive the model from a theory, for example by using differential or difference equations. The model derived is then tested by fitting it to data that are relevant to the property to be modelled. The second approach is largely atheoretical. One

* I am grateful to Gabriel Altmann (Bochum) and Manfred Lösch (Bochum) for their helpful comments on an earlier version of this article.

uses a set of data describing the property to be modelled to choose a model which in view of its mathematical properties is likely to fit the data. The model is then statistically fitted to the data set (sometimes to a different, but comparable set). In practice the two approaches are often combined.

In both approaches models are fitted to empirical data. This involves the estimation of parameters, the testing of statistical significance of the regression equation obtained and of individual parameters, and the measurement of both goodness of fit and predictive power of the model.

One can differentiate the following two basic types of model used in quantitative linguistics: (a) regression models, and (b) probability distributions. In this article the focus will be on regression models. Probability distributions will only be touched upon; they are discussed in some detail in Grotjahn and Altmann (in press).

The following subtypes of regression models might be distinguished: (a) functional models, (b) control models, and (c) predictive models (cf. Draper, Smith 1981: 412ff.).

If the true functional relationship between a dependent variable and a set of independent variables (the so-called predictors) is known for a specific problem, we are able to understand, and often also to control and predict the dependent variable. However, in practice it is often not possible to construct true functional models, and even if a functional model can be determined, it is usually nonlinear, and in addition often very complicated and difficult to interpret.

Moreover, a functional model is not always suited to control an independent variable for technological applications such as text generation or speech production. The model may contain variables which, although important, are difficult to measure or which cannot be measured at all in a given research situation (e.g., mental processes in text production). Sometimes experiments might be designed to assess the strength of the influence of a specific variable to be included in a control model. But this is rarely done in quantitative linguistics.

When a functional model is very complex and/or when important variables cannot (readily) be measured, a linear approximation can often be obtained, which, although unrealistic in some respects, can be used to

predict the dependent variable. Moreover, in such a predictive model, variables which are important for control might be omitted because they are highly correlated with variables included in the equation and thus are not needed for prediction (cf. Draper, Smith, 1981: 421f.). A predictive model may also provide hints for further research and help to choose among possible predictor variables.

In this article I shall not deal further with the various objectives which may be involved in model construction. Instead, I will discuss some potential pitfalls involved in the *evaluation* of regression models. Note that the criticism expressed throughout this article does not imply that some methods lead to models which are definitely incorrect, whereas other methods lead to models which are definitely correct. Instead, I would rather subscribe to the idea that "*all models are wrong*; some, though, are better than others and we can search for the better ones" (McCullagh, Nelder 1983: 6).

I shall first deal in some detail with the use of inferential statistics in regression analysis. The focus will be on the overall F test, which is the most common significance test for a regression model. The problems discussed include: large sample sizes, statistical significance vs. practical significance, violations of assumptions, nonlinear models, replicated responses.

In the next section the coefficient of determination R^2 is discussed at some length. The problems dealt with include: the choice of an appropriate coefficient, nonlinear models, pitfalls in the interpretation of R^2, and R^2 in the case of replicated responses. Subsequently some guidelines are proposed for the use of R^2.

The article concludes with a brief discussion of how to check the stability of parameters over the sample space (with a focus on the procedure of cross-validation) and with some suggestions as to how the construction and evaluation of regression models could be improved in quantitative linguistics as well as in other empirical sciences.[1]

[1] Although this article focusses on quantitative linguistics, most of what will be said also applies to other empirical disciplines such as (educational) psychology or sociology.

2. On the Use of Inferential Statistics

2.1 The Problem of Large Sample Sizes

One problem typical of quantitative linguistics results from the large sample sizes we are usually confronted with. There are at least two reasons why we often have large samples in linguistics: First, linguistic data are often more easily available than, for example, psychological data. Second, with regard to many linguistic phenomena representativeness can only be achieved by using very large samples.

With large sample sizes statistical tests tend to become significant even if the data correspond very well to the hypothesis being tested. To give an example: As the sample size n increases, the well-known t test for two means tends to become significant, even if the observed difference between the two sample means is minimal. In spite of this problem there are still quantitative linguists who rely primarily on significance tests in analyzing huge data sets (an example is discussed in Altmann 1992).

The problem of large sample sizes is a general problem of inferential statistics, and pertains to the evaluation of both regression models and probability distributions. The most common procedure to examine the goodness of fit of probability distributions is Pearson's chi-square test. The statistic used in this test belongs to a whole family of so-called power divergence statistics, which also includes the Freeman-Tukey statistic, the (modified) log likelihood ratio statistic and the Neyman modified chi-square statistic. With tests of this kind the following problem arises: if the difference between the theoretical and empirical frequencies is taken to be fixed, the chi-square statistic as well as the other power divergence statistics increase linearly with the increase of the sample size. This has the undesirable consequence that even minimal differences between model and reality lead to a highly significant value of the test statistic and thus to the possible rejection of the model (cf. Grotjahn, Altmann in press, who also discuss possible solutions to this problem.)

When we test the adequacy of a regression model with the usual overall F test for linear regression or when we evaluate the significance of individual parameters of a regression equation with the t test or the partial F test, a similar problem arises: the test tends to become significant

when the sample size increases. However, in contrast to goodness-of-fit tests of probability distributions a significant F or t test does not mean that the regression model is to be rejected, but rather that it is to be *accepted*. With large samples the usual F or t tests might thus lead to the acceptance of a regression model although its fit and predictive power is actually very poor.

2.2 The Classical Normal Linear Regression Model

To illustrate a number of problems in connection with the evaluation of regression models and in particular with regard to the F test, I shall first describe the classical normal linear regression model. It will be made clear that the model involves a number of important assumptions. Problems arising from violations of assumptions will be discussed in Section 2.4.

The model involves the following basic regression equation:

$$y_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} + \varepsilon_i. \qquad (1)$$

In equation (1) y denotes the regressand, that is, the dependent variable, which is considered to be some function of the regressors or independent (predictor) variables $x_1, x_2, \ldots, x_J$, the set of parameters $\{\beta_j\}$ and a random disturbance or error term ε. The subscript i $(i = 1, 2, \ldots, n)$ refers to the ith observation. Note that due to ε the relation expressed by equation (1) is stochastic (i.e., ε disturbs an otherwise deterministic relation).

To complete the specification of the classical normal regression model, Kmenta (1971: 348) adds seven *basic assumptions*, which are taken to apply to all observations (cf. also Kmenta 1971: 202; for a succinct exposition of the assumptions in matrix notation cf. Kmenta 1971: 350):

(A1) ε_i is normally distributed.

(A2) $E(\varepsilon_i) = 0$.

(A3) $E(\varepsilon_i^2) = \sigma^2$.

(A4) $E(\varepsilon_i \varepsilon_k) = 0$. $(i \neq k)$

(A5) Each of the explanatory variables is nonstochastic with values fixed in repeated samples and such that, for any sample size,

$$\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$$

is a finite number different from zero for every $j = 1, 2, \ldots, J$.

(A6) The number of observations exceeds the number of coefficients to be estimated, that is, $n > J$.

(A7) No exact linear relation exists between any of the explanatory variables.

The first two assumptions state that the vector of disturbances follows a multivariate normal distribution $N(\mathbf{0}, \mathbf{\Sigma})$. Since y_i is merely a linear function of ε_i, assumptions (A1) and (A2) imply that y_i is a normally distributed random variable.

The third assumption concerns *homoskedasticity* and means that every disturbance has the same variance σ^2. (A3) rules out, for example, that the variance of ε_i might be greater for higher than for lower values of x_{ij}.

The fourth assumption concerns *nonautoregression*. Assumptions (A2) and (A4) together imply that $Cov(\varepsilon_i, \varepsilon_k) = 0$ for all $i \neq k$. This means that the disturbance occurring at one point of observation must not be correlated with any other disturbance.

The fifth assumption requires that the values of the predictor variables be either controlled or fully predicted and that their values be the same from sample to sample. Note that this assumption rules out the possibility of considering the predictors to be random variables. The requirement that $(1/n)\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$ implies that the values of x_j in the sample must not all be equal, and should not increase or decrease without limit as the sample size grows.

Assumptions (A6) and (A7) are important for estimation. Assumption (A6) requires that there be a sufficient number of degrees of freedom for estimation. Assumption (A7) stipulates that none of the predictor variables be perfectly correlated with any other predictor variable or

with any linear combination of predictor variables (cf. the discussion of multicollinearity in Section 2.4.5 below).

The *method of least squares* is the most common procedure to estimate the parameters of the classical normal regression model. Least squares estimation consists in minimizing the squared deviations of the observed values from the predicted values. Under the assumption of the classical normal regression model the least squares estimators are equivalent to the *best linear unbiased estimators* and the *maximum likelihood estimators*. The least squares estimators thus have all the desirable properties: they are *unbiased* (i.e., their expectation corresponds to the population parameter to be estimated), *efficient* (i.e., they have the smallest variance of all unbiased (linear) estimators of the respective parameter), and, if certain regularity conditions are satisfied, they are also *consistent* (i.e., their sampling distribution tends to become concentrated on the true value of the parameter as sample size increases to infinity (cf. Kmenta 1971: 154ff: 205ff., 347ff.).

There are several generalizations of the classical normal linear regression model. For example, dropping the assumptions of homoskedasticity and nonautoregression leads to the generalized linear regression model (cf. Heil 1987, chap. 8; Kmenta 1971, chap. 12; Judge, Hill, Griffiths, Lütkepohl, Lee 1982: 247ff.). If the distribution of the response variable y is allowed to come from an exponential family (including, in addition to the normal distribution, e.g., the Poisson and the binomial distributions), we are dealing with generalized linear models as described by McCullagh and Nelder (1983). In this article, the discussion will be restricted to the classical normal regression model and to nonlinear regression, the latter being only briefly dealt with.

2.3 The *F* Test for Overall Regression

The F test for overall regression is based on the ratio of explained variance to unexplained variance. Let us designate the total sum of squares as

$$SST = \sum_{i=1}^{n}(y_i-\bar{y})^2, \qquad (2)$$

the sum of squares due to regression as

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{3}$$

and the residual sum of squares (i.e., the error sum of squares) as

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{4}$$

where $\hat{y}_i$ denotes the fitted value of y calculated on the basis of the set of estimated parameters $\{b_j\}$ and the values x_{ij} of the $x_j (j = 1, 2, \ldots, J)$ for $i = 1, 2, \ldots, n$.

If the parameters of the regression equation (1) have been estimated by using the method of least squares, then

$$SST = SSR + SSE. \tag{5}$$

The decomposition of SST into SSR and SSE crucially depends on how the regression parameters are estimated. With an estimation method that leads to a different sample regression line, equation (5) is not valid (cf. Kmenta 1971: 234). Since under the assumption of the classical normal linear regression model, the least squares estimators are equivalent to both the best linear unbiased estimators and the maximum likelihood estimators (cf. Kmenta 1971, chap. 7.3), only least squares estimation will be dealt with in this article.

The F statistic can now be expressed as follows (cf. Daniel, Wood 1980: 345):

$$F = \frac{SSR/df_1}{SSE/df_2} \tag{6}$$

with $df_1 = J$, $df_2 = n - J - 1$, J designating the total number of predictor variables and n the total number of observations.

The proportion of variance explained by the regression may be measured by the coefficient of determination R^2, which in the case of multiple linear regression corresponds to the squared multiple correlation coefficient between the predictor variables and the dependent variable (cf. Section 3). The coefficient of determination may be defined as

$$R^2 = \frac{SSR}{SST} \tag{7}$$

or using (5) as

$$R^2 = 1 - \frac{SSE}{SST}. \tag{8}$$

With the help of R^2 the F statistic may also be expressed as follows:

$$F = \frac{R^2}{1 - R^2} \frac{(n - J - 1)}{J}. \tag{9}$$

The F statistic tests the null hypothesis that the multiple correlation between the predictor variables and the dependent variable is zero in the population. This corresponds to testing the null hypothesis H_0: that all the β's (excluding β_0) are zero against the alternative hypothesis H_1: that at least one of the β's (excluding β_0) is not zero (cf. Draper, Smith 1981: 94). H_0 is rejected if $F > F_{J,n-J-1;1-\alpha}$.

Note that strictly speaking the test of the hypothesis that the multiple correlation between the predictor variables and the dependent variable is 0 in the population is not completely equivalent to the test of the hypothesis that the β's are 0. In the latter case, the predictor variables are considered as nonstochastic, in the first case as stochastic (cf. Helland 1987, Sampson 1974 and Section 2.4.4 below). As has been shown by Sampson (1974), the rejection regions and the statistics for the F test are the same for both models, while the corresponding power functions differ.

2.4 Violation of Assumptions

The F test is based on the assumption that the classical normal linear regression model as specified in Section 2.2 is valid and that furthermore the method of least squares (or maximum likelihood or best linear unbiased estimation) is used to obtain the parameters. The same holds for other common significance tests used in regression analysis (e.g., the t test for individual regression coefficients) and for the construction of confidence intervals. Violation of one or several of the assumptions on which the classical normal regression model is based, or the use of a method of estimation other than least squares (or maximum likelihood

or best linear unbiased estimation) may severely bias the results of inferential statistics (cf. Draper, Smith, 1981: 87; Kmenta 1971: 247f.).

In quantitative linguistics the various assumptions involved in regression analysis are as a rule neither explicitely stated nor are the models checked for possible violations of assumptions. The assumptions appear either not to be known or to be considered not very important. Even in elementary and intermediate texts on applied regression analysis the various assumptions are usually discussed only very briefly, and often important aspects receive a very perfunctory treatment or even none at all. A common argument for the neglect of assumptions, found in a standard text on applied regression analysis, reads as follows:

> It has convincingly been shown that the F and t tests are "strong" or "robust" statistics ... In general it is safe to say that we can ordinarily go ahead with analysis of variance and multiple regression analysis without worrying too much about assumptions. (Kerlinger, Pedhazur 1973: 47f.)

It will be shown that we do have to worry about assumptions, since violations of assumptions may have various effects, some of which even disastrous, on the validity of the classical normal regression model and more specifically on the validity of the F test (and t test) (cf. also Tabachnick, Fidell 1989: 71).

Some consequences resulting from violations of the assumptions outlined in Section 2.2 will now be briefly discussed. Thereafter, two problems in connection with the functional form of the basic regression equation (1) of the classical normal regression model will be dealt with: (a) the problem of an incorrect specification of the regression equation as to the number of explanatory variables, and (b) problems arising in the case of nonlinear models. The first problem will be touched upon at the end of this section; the second problem will be dealt with in some detail in Section 2.6.

2.4.1 Nonnormality of the Disturbance

If the assumption of normality of the disturbance is dropped, the least squares estimators are still unbiased, have the smallest variance of all linear unbiased estimators and, if certain regularity conditions are satisfied, they are also consistent. However, they are no longer efficient and no longer maximum likelihood estimators. Note that if we know

that the errors follow a specific distribution, for example, the highly leptokurtic double exponential distribution, the method of maximum likelihood should be used for parameter estimation rather than that of least squares (cf. Box, Draper 1987: 83f.). Moreover, without the assumption of normality of errors, the least squares estimators of the regression coefficients are not normally distributed in *small* samples, and significance tests and confidence intervals may be biased. However, according to the central limit theorem, the least squares estimators are asymptotically normally distributed as $n \to \infty$ (cf. Kmenta 1971: 248 and Heil 1987: 268). Therefore, when the sample size is large, as is often the case in quantitative linguistics, the assumption of the normality of errors does not appear to be crucial.

2.4.2 Heteroskedasticity

It may, for example, happen that some of the observations are less reliable than others and that thus the variance of the disturbance ε varies. By the same token, there may be different degrees of variation for different values of a predictor variable due to a priori restrictions.

In the case of heteroskedasticity the ordinary least squares estimation method does not apply and significance tests and confidence intervals based on this method are wrong (cf. Kmenta 1971: 255). Yet, when the variance changes in a known way, the method of *weighted* least squares can be used (cf., e.g., Draper, Smith 1981: 108ff.).

There is evidence that in quantitative linguistics the assumption of homoskedasticity is quite often not fulfilled (cf., e.g., the data presented in Köhler 1986 or Hammerl 1991). Nevertheless, homoskedasticity is as a rule assumed and the problems involved seem to be widely ignored. One of the few exceptions is Hammerl (1991: 156) who states that in his investigations the requirement of homoskedasticity is not always met. This is one reason why he rejects the use of the F test and instead favours the application of the coefficient of multiple and partial determination (cf. also Hammerl 1991: 31).

2.4.3 Autoregressive Disturbances

The assumption of nonautoregression is often violated in time series data, in particular if the interval between the consecutive observations is short (cf. Kmenta 1971: 269-297). By the same token, the assump-

tion may be violated when we sample from a continuous stretch of text and the distances between the observations are short.

Suppose we want to predict word length by means of clause length, and sample consecutive clauses of a text to this end. We know that clause length is in turn affected by sentence length. If sentence length is not included in the regression equation, it will affect the prediction as disturbance. As a result the disturbance occurring in two consecutive clauses of one and the same sentence might be correlated.

When the disturbances are autoregressive, the conventional least squares estimators of the regression coefficients are no longer (asymptotically) efficient and the estimates of their variances are biased. Biasedness of the estimates of the variances implies that "the conventional formulas for carrying out tests of significance or constructing confidence intervals with respect to the regression coefficients lead to incorrect statements" (Kmenta 1971: 281). More precisely, especially when autocorrelation is positive, we may considerably underestimate the variances of the regression coefficients and thus overestimate the quality of the empirical regression coefficients (cf. Schönfeld, 1969: 140). As a consequence, acceptance regions and confidence intervals will often be *narrower* than they should be for the specified α-level (cf. Kmenta 1971: 282).

Krämer, Kiviet and Breitung (1990) have shown that the F test becomes extremely non-robust when the autocorrelation among the disturbances increases. More specifically, the true rejection probability may be several times as large as the nominal rejection probability. This is in particular true for large samples. In small samples, the difference between the nominal and the true rejection probabilities is less pronounced, but still quite large for all practical purposes. Note that in small samples the difference is larger for non-intercept models (cf. Table 1 in Krämer, Kiviet and Breitung 1990). When using the F test in the case of positively autocorrelated disturbances, we thus may run a high risk of erroneously rejecting the null hypothesis.

There are various statistical procedures to test for the absence of autoregression (for a survey cf. King 1987). The most common procedure is the Durbin-Watson test, which is implemented in all major statistical packages. For certain design matrices, however, the power of the

Durbin-Watson test as well as that of many other autocorrelation tests may drop to zero as the autocorrelation among the disturbances increases (cf. Krämer, Zeisel 1990).

If statistical testing indicates autoregression, one might re-estimate the equation, using a method especially suited for autoregression (cf., e.g., Dillon, Goldstein 1984 and the references given therein). Alternatively, one might re-examine the regression model since autoregression of the disturbance may be the result of some unexplained systematic influence (such as sentence length in the above example).

2.4.4 Stochastic Predictors

Assumption (A5) requires that each of the predictor variables is fixed in repeated samples. If we allow the values of the predictors to occur with certain probabilities, we are dealing with stochastic predictors. Stochastic predictors are quite common in quantitative linguistics.[2] The regression models actually used do, however, assume that the predictors are nonstochastic. Potential problems resulting from this approach appear to be ignored.

If a predictor is stochastic but independent of the disturbance, the least squares estimators of the regression coefficients retain all their desirable asymptotic properties (cf. Kmenta 1971: 297ff.). Serious problems may, however, arise if the stochastic predictor and the disturbance are correlated as in the distributed lag model or in errors-in-variables models. In this case the results of ordinary least squares estimation do not hold even asymptotically. An explanation for this fact is that if the predictor variable and the disturbance are correlated, the decomposition of SST into SSR and SSE is no longer valid since it does not take into account the *joint* effect of x and ε on y (cf. Kmenta 1971: 303).

If the predictor variable and the disturbance are correlated, more complicated fitting procedures than the method of ordinary least squares

[2] Stochasticity of the predictor variables is, for example, assumed if the inverse regression is calculated (cf., e.g., Hammerl 1991: 91 who exlicitely takes into account the possibiliy of interchanging the dependent and independent variables in his models). If the independent variables were in fact nonstochastic, they could not become a stochastic dependent variable in the inverse regression.

must be used. According to Draper and Smith (1981: 124), the random variation of a stochastic predictor may, however, be safely ignored and the method of ordinary least squares used if we can assume that the random variation is small compared to the observed range of the variable.

2.4.5 Multicollinearity

Assumption (A7) requires that none of the predictor variables is perfectly correlated with another predictor variable or with a linear combination of other predictor variables. When this assumption is violated, we are dealing with *perfect multicollinearity* (cf. Kmenta 1971: 380ff.).

In the case of more than two predictor variables, perfect correlation between two predictor variables is only a sufficient but not necessary condition for perfect multicollinearity. In the case of more than two predictor variables, perfect multicollinearity may arise even if none of the correlation coefficients is particularly high. This means that as a rule we cannot simply look at the coefficients of correlations or at the plots of the x_j against y to conclude that there is perfect or high multicollinearity (cf. Kmenta 1971: 382ff.; Daniel, Wood 1980: 50-53 and the discussion of suppressor variables in Section 3.4 below).

When perfect multicollinearity obtains, the sum-of-squares-and-cross-products matrix becomes singular and its inverse does not exist. This means that the solution of the normal equations and the least squares estimators of the regression coefficients are nonunique.

When at least one of the predictor variables is highly correlated with another predictor variable or with a linear combination of other predictor variables, we speak of *high multicollinearity*. In the case of high multicollinearity, the solution of the normal equations and hence the estimates of the regression coefficients are unique, but tend to be extremely inaccurate. One (controversial) way of coping with this situation is to use the ridge regression technique, which is available, for example, as part of the well-known BMDP software. Good expositions of ridge regression can be found in Marquart and Snee (1975) or Dillon and Goldstein (1984, chap. 7).

In quantitative linguistics the problem of multicollinearity may arise when one attempts to model the intricate interrelationships among the numerous pertinent variables with the help of complex path models.

The higher the degree of multicollinearity, the larger the variances and covariances of the least squares estimators of the regression coefficients (measures of multicollinearity are discussed, e.g., in Willan and Watts 1978). This entails among other things that the confidence interval for a given regression coefficient will be wide and the power of the corresponding significance test weak (cf. Kmenta 1971: 391). This should be kept in mind when using inferential techniques.

2.4.5 Specification Errors

Another assumption implicitely involved in ordinary least squares estimation is that the regression equation is correctly specified with regard to the number of relevant explantory variables. Kmenta (1971: 394) has pointed out that the *omission* of relevant explanatory variables may lead to biased tests of significance and biased confidence intervals. However, if an irrelevant explanatory variable is included, the usual methods of inference remain valid (cf. Kmenta 1971: 399; cf. also Binkley, Abbott 1987; Rao, Miller 1971: 67f.; Pindyck, Rubinfeld 1981: 262). Note that the kind and magnitude of bias depends on whether the regressors are considered to be fixed or stochastic (cf. Binkley, Abbott 1987 and Kinal, Lahiri 1983).

2.5 Interpretation of the F Test

From formula (9) above it can easily be seen that for fixed values of R^2 and J, F increases when the sample size n increases. F thus becomes necessarily significant with large sample sizes. Consequently, the test of the null hypothesis "becomes trivial because it is almost certain to be rejected" (Tabachnick, Fidell 1989: 154).

Furthermore, a significant overall F value only means that given a specified α level the proportion of the variation observed in the data and accounted for by the regressor variables is greater than would be expected by chance in $100(1-\alpha)\%$ of similar sets of data with the same values for n and the regressor variables. This does not, however, necessarily imply that the equation obtained is also useful for prediction. It will be useful for predictive purposes only if the amount of variation accounted for by the fitted equation is also *substantial*, that is, when it is large compared to the random error. Hence a fitted equation will often be of no value even though a highly significant F value has been obtained (cf. Draper, Smith 1981: 93 and Cohen 1977, chap. 9). This

is one reason why the F test is seldom used in applied econometrics (cf. Heil 1987: 221).

Work by Wetz (1964) suggests that for a "useful" as distinct from a "significant" regression equation, it is not sufficient that the observed F value merely exceeds the selected percentage point of the F distribution, but that it should exceed the selected percentage point by a *multiple* (cf. Draper, Smith 1981: 93, 129-133, cf. also Draper 1984). The multiple can be determined by comparing the range of response values predicted by the regression equation obtained with the standard error of the responses. The exact value of the multiple to be chosen is to a great extent arbitrary just as the choice of a significance level is. Tables provided by Draper and Smith (1981: 132f.) suggest, however, that an observed F ratio must be at least four or five times the usual percentage points of the F distribution.

The problem that a model may be inadequate in spite of a highly significant overall F test seems not to be taken into account by a number of quantitative linguists who use the F test not only as a test for regression but – illegitimately – at the same time as a *measure of goodness of fit* (cf. Section 2.8 below).

One quantitative linguist who explicitely recognizes that a highly significant F value does not necessarily entail that a regression equation is also practically useful is Hammerl (1990, 1991). He states with regard to his data that the F test was always significant although the coefficient of determination R^2 indicated that the amount of variance explained was relatively small (cf. Hammerl 1990: 9).

A further complication in the interpretation of the F test arises in the case of two or more predictor variables. When there are several predictors, it is quite possible that the F test turns out to be significant although none of the β's (excluding β_0) is significantly different from zero according to the t test. Kmenta (1971) interprets this situation as follows:

> In this case, we would *reject* the hypothesis that there is no relationship between y on the one side and $x_1, x_2, \ldots, x_J$ on the other side, but we would *not reject* the hypothesis that any one of the explanatory variables is irrelevant in influencing y. Such a situation indicates that the separate influence of each of the explanatory variables is weak relative to their joint influence on y. This is symptomatic of a high

> degree of multicollinearity, which prevents us from disentangling the separate influences of the explanatory variables. (p. 390)

2.6 Nonlinear Models

2.6.1 Nonlinearity in the Variables and Nonlinearity in the Parameters

Nonlinear models are extensively used in quantitative linguistics. When we say that a model is nonlinear, we may refer to nonlinearity in the variables and/or nonlinearity in the parameters to be estimated. Nonlinearity in the parameters means in the context of this section that when we estimate the parameters, they are involved in a nonlinear way (cf. Draper, Smith 1981: p. 459). For example, the function

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \tag{10}$$

is nonlinear in the variables, but linear in the parameters, whereas the function

$$y = \exp(\beta_1 + \beta_2 x^2 + \varepsilon) \tag{11}$$

is nonlinear in the parameters as well as in the variables (cf. Draper, Smith 1981: 218ff., 458ff.).

In this article, only models which are nonlinear in the parameters will be referred to as nonlinear. Such models are often used in quantitative linguistics and particularly in synergetic linguistics (cf., e.g., Altmann 1983; Altmann, von Buttlar, Rott, Strauß 1983; Hammerl 1991; Köhler 1986; Zörnig, Köhler, Brinkmöller 1990).

It will be shown that the F test is as a rule biased in the case of nonlinear models. Draper and Smith (1981: 484) point out that this is also true for other common tests of significance. The authors state: "The usual tests which are appropriate in the linear model case are, in general, *not* appropriate when the model is nonlinear." This fact should be taken into account when interpreting work with nonlinear models in quantitative linguistics.[3]

[3] For a brief discussion of (asymptotically valid) tests and confidence regions for nonlinear models see Amemiya (1983: 347-353) or Fahrmeier, Kaufmann and Kredler (1984: 149-153).

2.6.2 The Generalized Multiple Linear Regression Equation

Models that are nonlinear in the variables but linear in the parameters will be considered as *linear*. These models can be viewed as special cases of a linear (in the parameters) model of the following most general form (cf. Draper, Smith 1981: 218ff.):

$$y = \beta_0 z_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \varepsilon, \tag{12}$$

where z_0 is a dummy variable which is always unity, and each z_j $(j = 1, 2, \ldots, p)$ is a general function of $x_1, x_2, \ldots, x_J$.

From equation (12), we obtain with $p = 1$ and $z_1 = x$ the simple first-order regression model with one predictor variable. (The value of the highest power of a predictor variable in a regression model is called the *order* of the model.) If $p = J$ and $z_j = x_j$, we obtain the first-order multiple regression equation as defined by equation (1) in Section 2.2. With $p = 2, z_1 = x_1$ and $z_2 = x^2$, we obtain the second-order polynomial model given in equation (10) above. Note that polynomial models of any order can be represented by equation (12) by simply renaming the predictor variables.

There are many possible transformations, all leading to special cases of equation (12).[4] For example, if we use the logarithmic transformation $z = \ln x$ and the square root transformation, we obtain for $p = 2$ the models (13) and (14):

$$y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \varepsilon \tag{13}$$

$$y = \beta_0 + \beta_1 x_1^{1/2} + \beta_2 x_2^{1/2} + \varepsilon. \tag{14}$$

2.6.3 Intrinsical Linearity vs. Intrinsical Nonlinearity

We can distinguish between two basic types of *nonlinear* models (i.e., nonlinear in the parameters to be estimated): *intrinsically linear* and *intrinsically nonlinear* models. A nonlinear model is intrinsically linear when it can be converted to the standard linear model of equation

[4] References to the literature on linearizing transformations can be found, e.g., in Draper and Smith (1981: 683f.).

(12). If a model cannot be expressed in the form of equation (12), it is intrinsically nonlinear (for a different definition of intrinsically (non)linear models cf. Kmenta 1971: 451).

While equation (11) can be converted, by taking logarithms, into the form

$$\ln y = \beta_1 + \beta_2 x^2 + \varepsilon, \tag{15}$$

which is the form of equation (12) and is linear in the parameters, it is impossible to transform the following equation into a form linear in the parameters (cf. Draper, Smith 1981: 458f.):

$$y = \frac{\beta_1}{\beta_1 - \beta_2}(e^{-\beta_2 x} - e^{-\beta_1 x}) + \varepsilon. \tag{16}$$

In quantitative linguistics, Hammerl's (1991) or Köhler's (1986) various nonlinear synergetic models are all intrinsically linear, whereas the differential equation models proposed by Zörnig, Köhler and Brinkmöller (1990) for the oscillation of word length as a function of word frequency are intrinsically nonlinear.

2.6.4 Intrinsically Linear Models

Any model that can be written, perhaps after a suitable transformation, in the form of equation (12) can be analyzed by the general methods developped for the classical normal regression model (cf. Draper, Smith 1981: 219). If the transformation involves only a relabelling of the predictor variables as in the case of the polynomial model in equation (10), which is nonlinear in the variables but linear in the parameters, the methods developed for the classical normal linear regression model apply without any modification (cf. Kmenta 1971: 451). To estimate the parameters of equation (10), for example, we only need to transform it into a multiple linear regression equation with two predictor variables as indicated above. We then apply the same transformation to our data and estimate the three parameters of the linearized model in the usual way. Note, however, that the z's will often be highly correlated, so that the variances of the parameter estimates may be large (cf. Kmenta 1971: 452).

However, if the linearization of nonlinear models involves the transformation of the parameters to be estimated, a number of problems arise

which are usually not taken into account in quantitative linguistics. For example, when we transform, by taking logarithms, the function

$$y = \alpha x^{\beta} e^{\varepsilon}, \tag{16}$$

into

$$\ln y = \ln \alpha + \beta \ln x + \varepsilon, \tag{17}$$

the linearization involves the transformation of the parameter α. If we then use the method of least squares to estimate α, the estimate is a valid least squares estimate only with regard to the linearized function. With regard to the original nonlinear function, however, the estimate of α does not possess the small-sample properties of the least squares estimator. It is, in particular, biased and the amount of bias cannot be exactly determined (cf. Schneeweiß 1974: 52 and Kmenta 1971: 458). As a consequence, the F test for overall regression is only valid with regard to the linearized model, but biased with regard to the nonlinear model.

The same problem arises if, for example, a power function with an intercept or an exponential function with or without an intercept is linearized with the help of a logarithmic transformation. (These models are extensively used in quantitative linguistics and they are usually linearized for parameter estimation.) In all these cases the estimate of the linearized regression equation is biased with regard to the nonlinear function and thus not a least squares estimate of the nonlinear function (cf. Draper, Smith 1981: 224; Rützel 1974).

Another nonlinear function whose linearization by means of a logarithmic transformation results in a biased least squares estimator for the nonlinear model is Altmann's (1983) "law of change", which is a special case of a nonlinear growth model. The law of change reads as follows (note that there is no disturbance term and that the parameters to be estimated are a and k):

$$p = \frac{1}{1 + ae^{-kt}}. \tag{18}$$

Linearization yields

$$\ln\left(\frac{1}{p} - 1\right) = \ln a - kt. \tag{19}$$

Again, the parameter a is transformed when the model is linearized. Consequently, the least squares estimate as well as the F test carried out by Altmann (1983: 111) are biased with regard to the nonlinear model (cf. also Altmann, von Buttlar, Rott, Strauß 1983: 62)

Nonlinear growth models involve the further problem that growth data do not always satisfy the assumptions on which ordinary least squares estimation is based. For example, the observations may be correlated or the errors may be heteroskedastic (cf. Draper, Smith 1981: 506). This is very often not taken into account.

More adequate least squares estimates of the parameters of nonlinear functions can be obtained if we use the *method of nonlinear least squares*. With this method the principle of least squares is *directly* applied to the *nonlinear* function. This implies that the parameters are estimated so that the squared deviations from the *nonlinear* function is a minimum. There are, however, a number of technical problems connected to this method. The minimization procedure results as a rule in normal equations which must be solved iteratively with the help of a computer. In some cases it is very easy to develop an iterative technique for solving the equations. However, if more complicated models are fitted, the solution of the normal equations may be very difficult and it may even happen that multiple solutions exist (cf. Draper, Smith 1981: 462). For this reason it is often advisable to use an approximation such as the method of linearization as described above or the Taylor series expansion as described below.[5]

In quantitative linguistics, the parameters of nonlinear models which can be linearized are, as a rule, not estimated by means of nonlinear least squares, but by applying the method of least squares to the

[5] An introduction to nonlinear estimation can be found, e.g., in Draper and Smith (1981, chap. 10). Nonlinear estimation is implemented, e.g., in the BMDP software. For references to other programs for nonlinear regression analysis cf., e.g., Daniel, Wood (1980) and Draper, Smith (1981, chap. 10). Formulas for the nonlinear least squares estimators of the power and exponential functions are given by Rützel (1974).

linearized version of the nonlinear model. This is the procedure, for example, in Köhler (1986) and Hammerl (1991) who both use, among other models, power functions of the type of equation (16). As a consequence both the parameter estimates and the corresponding F tests for overall regression are biased (the F test is calculated only by Köhler).

Furthermore, the interpretation of the coefficient of determination, which Hammerl (1991) uses for model evaluation, also becomes problematic. If the parameter estimates are not least squares estimates, the decomposition of SST into SSR and SSE is no longer valid and the value of the coefficient of determination can no longer be interpreted as the proportion of variance explained (cf. also Section 3.4 below). These problems appear to be ignored not only by Köhler and Hammerl, but also by most other quantitative linguists.[6]

The above criticism does not, of course, imply that the models obtained through linearization are worthless. Linearization is in many cases a very convenient method which yields a good approximation to the original nonlinear model as fitted with the help of more sophisticated methods. In some cases, however, the bias may be considerable. This should be kept in mind when the method of linearization is used or when a nonlinear model fitted after linearization is evaluated.

2.6.5 The Taylor Series Expansion

Another widely used approximative method for the estimation of the parameters of a nonlinear model involves a linearization of the nonlinear function by means of a Taylor series expansion. This method can be used not only for nonlinear models that are intrinsically linear but also for models that are intrinsically nonlinear. The reason for its general applicability is that any function that has a continuous pth derivative can be written as a Taylor series expansion.

Suppose we already have initial values for the parameters to be estimated. These may be, for example, rational guesses, or preliminary estimates obtained by short-cut methods of fitting such as those described in Altmann (1983) for nonlinear growth models. If we carry out

[6] The fact that the F test is biased in the case of nonlinear models is explicitely acknowledged by Hammerl (1990: 8) and Hammerl (1991: 156).

a Taylor expansion of the function about the vector of initial values, and truncate the expansion at the first derivatives, we obtain a linear approximation to the original nonlinear function (cf. Draper, Smith 1981: 462ff.). The parameters of the linearized function are then estimated in the usual way. Subsequently the estimates are taken as new initial values and the entire process (i.e., expansion in a Taylor series about the new initial values and calculation of new estimates) is started again. This iterative process is continued until a predefined degree of precision is obtained.

However, this method has possible drawbacks as well: It may converge very slowly, that is, a very large number of iterations may be required. It may oscillate widely, and it may even not converge at all (cf. Draper, Smith 1981: 464). Furthermore, since the higher-order terms of the Taylor series are truncated, the approximation of the linear function to the true model may be poor and the estimates severely biased (cf. also Kmenta 1971: 399f.). As a consequence, the results of the F test for overall regression and of other significance tests may be severely biased as well.

In quantitative linguistics, the method of fitting a nonlinear model by means of an expansion into a Taylor series has been used, for example, by Altmann (1983) or Altmann and Kind (1983). In both articles, the parameters obtained have been tested for significance. In the light of what has been pointed out, the results of the corresponding tests should be viewed with caution.

2.6.6 Specification of the Disturbance Term

A final problem to be discussed in the context of nonlinear models concerns the specification of the disturbance term. In quantitative linguistics, just as in economic theory, all relationships are stated as a rule in a deterministic form, that is, without an error term. This means that when a nonlinear model is derived on the basis of a set of theoretical assumptions about the nature of the domain to be modelled (leading, e.g., to a system of differential equations), the existence of a random disturbance is usually not taken into account either theoretically or mathematically, possibly for reasons of (mathematical) simplicity and practicability (cf., e.g., the examples in Altmann 1983; Altmann, von Buttlar, Rott, Strauß 1983; Hammerl 1991; Köhler 1986).

This does not, however, imply that the possible influence of factors not accounted for in the model is not taken into consideration. Hammerl (1991: 91), for example, when constructing his lexical models, explicitely assumes that the influence of all factors not accounted for by the model is *constant* over the entire range of the dependent variable. A similar assumption is made by Köhler (1986: 98). I think that this is not a very realistic view. It would be more appropriate to consider the influence of the factors not accounted for to be random, and then model the random influence in the form of a disturbance term explicitely specified in the corresponding nonlinear (or linear) model. Admittedly, this would complicate the task of theory building and mathematical modelling.

The fact that the existence of random disturbance is not explicitely taken into accout when a (nonlinear) model is constructed, does not, however, mean that the model is considered to be deterministic. (If this were actually the case, the application of inferential statistics, e.g. in the form of the F test, would make no sense.) Rather, the disturbance term enters at a later stage of model construction, namely when the nonlinear model is linearized (which is the most common procedure) and is treated as a classical normal regression model with an additive disturbance.

The latter approach involves, however, a number of potential pitfalls. For example, suppose we have a multiplicative model of the type

$$y = \alpha x_1^{\beta} x_2^{\gamma} x_3^{\delta} \varepsilon, \tag{20}$$

where α, β, γ and δ are the parameters to be estimated and ε is a multiplicative random error. Taking logarithms converts the model into the linear form

$$\ln y = \ln \alpha + \beta \ln x_1 + \gamma \ln x_2 + \delta \ln x_3 + \ln \varepsilon. \tag{21}$$

In the linearized function, the disturbance term thus enters as $\ln \varepsilon$ rather than ε. This entails that tests of significance and confidence intervals are only valid if the logarithm of the disturbances rather than the disturbances themselves follow a multivariate normal distribution $N(\mathbf{0}, \mathbf{\Sigma})$ (cf. Draper, Smith 1981: 222f.) This and other potential problems become evident only if the disturbance term is explicity specified in the nonlinear model.

2.7 The F Test in the Case of Replicated Responses

It has been shown that the overall F test *presupposes* that the model to be tested is correctly specified with regard to its functional form and the number of relevant predictor variables. When there is lack of fit because the functional form of the equation is not correct (e.g., nonlinear instead of linear), the overall F test is biased.

If there are replicated responses, lack of fit can be tested. If this test turns out to be significant, the overall F test for regression is not valid. Since the problem of replicated responses appears to be widely ignored in quantitative linguistics, and since it is often not discussed in books on regression analysis, it will be dealt with in some detail (for the following cf. Box, Draper 1987: 70-74; Chang, Afifi 1987; Draper, Smith 1981: 35ff.).

Replicated responses are repeat mesurements of y. This means that two or more measurements have been made at the same value of x. For this reason, one also speaks of 'repeat x values'.

2.7.1 One Predictor Variable

Let us first consider the case of simple linear regression with one predictor variable x. Suppose we have the following five values for (x, y): $(1.3, 2.3)$, $(1.3, 1.8)$, $(2.0, 2.8)$, $(2.0, 1.5)$, $(2.7, 2.2)$. We say that there are two repeat observations at $x = 1.3$, two repeat observations at $x = 2.0$, and one repeat observation at $x = 2.7$. Note that a repeat measurement is only then a genuine replicated response if the measurements are made on *different* elements of the sample (they must not be repeated measurements of the same element).

The repeats can be used to estimate σ^2 (i.e., the variance of y). Such an estimate represents "pure error", which cannot be explained by the regression equation. This estimate of σ^2 is usually much more reliable than any other estimate.

Suppose there are K different values of x and, at the kth of these K values, $x_k (k = 1, 2, \ldots, K)$, there are n_k observations. Let y_{ku} be the uth observation $(u = 1, 2, \ldots, n_k)$ at x_k. Altogether, there are $N = \sum_{k=1}^{K} n_k$ observations (cf. Draper, Smith 1981: 35ff.; for a more rigorous treatment see Chang, Afifi 1987).

Pooling the internal sums of squares from all the replicated responses at the different x_k, we obtain the total pure error sum of squares, which corresponds to the within sum of squares (SSW) in analysis of variance:

$$SSW = \sum_{k=1}^{K} \sum_{u=1}^{n_k} (y_{ku} - \bar{y})^2 \tag{22}$$

with degress of freedom

$$df_W = \sum_{k=1}^{K} (n_k - 1) = N - K. \tag{23}$$

The pure error mean square, which corresponds to the within mean square (MSW) in analysis of variance, is then

$$MSW = \frac{SSW}{N - K}, \tag{24}$$

which is an estimate of σ^2 irrespective of whether the regression model being fitted is correct or not.

It can easily be shown that the pure error sum of squares is in turn part of the residual sum of squares. If we write the residual for the uth observation at x_k as

$$y_{ku} - \hat{y}_k = (y_{ku} - \bar{y}_k) - (\hat{y}_k - \bar{y}_k),$$

square both sides and sum over both u and k, we obtain

$$\sum_{k=1}^{K} \sum_{u=1}^{n_k} (y_{ku} - \hat{y}_k)^2 = \sum_{k=1}^{K} \sum_{u=1}^{n_k} (y_{ku} - \bar{y}_k)^2 - \sum_{k=1}^{K} (\hat{y}_k - \bar{y}_k)^2. \tag{25}$$

The left side of equation (25) is the error sum of squares (residual sum of squares) SSE and the first term on the right side is the pure error sum of squares SSW. The second term on the right side is called the lack of fit sum of squares (SSL). If there is one predictor variable, the residual sum of squares has $df_E = N - 2$ degrees of freedom. Hence, the number of degrees of freedom for lack of fit is $df_L = df_E - df_W = K - 2$, and the lack of fit mean square is $MSL = SSL/(K - 2)$.

There is perfect fit when the K predicted values are equal to the K means of the replicated responses. In this case, $SSL = 0$. Lack of fit can be tested by comparing the ratio $F = MSL/MSW$ with the $100(1 - \alpha)\%$ point of an F distribution with $K - 2$ and $N - K$ degrees of freedom. If the F value is *significant*, the model is considered to be inadequate, and the F for regression should *not* be calculated. If the F value is *not significant*, both pure error and lack of fit mean squares can be used as estimates of σ^2. We then recombine SSW and SSL into SSE and carry out the usual F test for regression. A numerical example for the use of the F test in the case of replicated responses can be found in Draper and Smith (1981: 38f.).

2.7.2 Several Predictor Variables

The formulas for one predictor variable are in principle also valid for the multiple predictor case. Note that in the case of multiple predictors a set of replicated responses must all have the same values for an explanatory vector $(x_1, x_2, \ldots, x_J)$. Further note that if there are replicated responses, the number, J, of predictor variables is limited by K, the number of distinct explanatory vectors, and not by N, the total number of observations (cf. Draper, Smith 1981: 42 and Chang, Afifi 1987: 195).

The complete ANOVA for lack of fit and for overall regression in the case of multiple linear regression with replicates is shown in Table 1. Note that the F test for regression is only valid if the F test for lack of fit is not significant.

Table 1. ANOVA for Replicated Responses

Source	df	SS	MS	F ratio
Regression	J	SSR	MSR	F = MSR/MSE
Residual	N-J-1	SSE	MSE	(test for regression)
Lack of fit	K-J-1	SSL	MSL	F = MSL/MSW
Pure error	N-K	SSW	MSW	(test for lack of fit)
Total	N-1	SST		

J = number of predictor variables
K = number of distinct explanatory vectors
N = total number of observations

The corresponding sums of squares are:

$$SSR = \sum_{k=1}^{K}(\hat{y}_k - \bar{y})^2 \tag{26}$$

$$SSE = \sum_{k=1}^{K}\sum_{u=1}^{n_k}(y_{ku} - \hat{y}_k)^2 \tag{27}$$

$$SSL = \sum_{k=1}^{K}(\hat{y}_k - \bar{y}_k)^2 \tag{28}$$

$$SSW = \sum_{k=1}^{K}\sum_{u=1}^{n_k}(y_{ku} - \bar{y}_k)^2 \tag{29}$$

$$SST = \sum_{k=1}^{K}\sum_{u=1}^{n_k}(y_{ku} - \bar{y})^2 \tag{30}$$

Like the F test for regression, the F test for lack of fit is subject to the problem of large sample sizes. We thus also need a *measure* for the goodness of fit of a regression model in the case of replicated responses. This problem will be dealt with in Section 3.5.

2.8 Some Further Examples from the Literature in Quantitative Linguistics

In this section I would like to illustrate some of the problems encountered by means of further examples taken from the literature in quantitative linguistics. The focus will again be on the F test. I would like to emphasize that the purpose of this section is not to single out some authors and criticize their work, but to point out (potential) pitfalls in the application of the F test.

In Köhler's (1986) pioneering work on synergetic linguistics, a number of regression models are derived and statistically evaluated. The author's approach to estimation and model validation is, however, inadequate for several reasons.

Köhler relies exclusively on the F test for the statistical evaluation of his models. Basing the evaluation solely on the F test is, however,

inappropriate. One should, for example, also carry out an analysis of the residuals, calculate indices for goodness of fit and predictive power (such as the coefficient of determination), and, if possible, cross-validate the regression equation. This criticism applies incidentally to the bulk of research in quantitative linguistics.

Moreover, Köhler's use of the F test is incorrect in several respects.

First, Köhler linearizes (as do many other authors) his nonlinear models with the help of a logarithmic transformation. Then the F test is used to evaluate the fit of his models. As has been shown in Section 2.6, the F test is biased in this case.

Second, there are a large number of replicated responses in Köhler's data (cf. the corresponding statement on page 98, footnote 55). These are dealt with by the author as follows: he calculates the means of the replicates and fits his model to the means without taking the number of replicates into account. The F test is then calculated with the help of equation (6) above, with n designating the number of means.

The same approach has been used by Hammerl (1991). However, in contrast to Köhler, Hammerl did not calculate the usual F statistic but the coefficient of determination R^2 to evaluate the fit of his models.

Hammerl's rationale for fitting his equations directly to the means of the replicates is as follows: he considers the means of the replicates, $\bar{y}_k$, as macro elements and the individual resonses, y_{ku}, as micro elements. Since the latter are affected by a large number of factors which are difficult to account for, Hammerl argues that calculating the means of the replicated responses will help to neutralize part of the effect of the random disturbance not accounted for. He further contends that in synergetic linguistics we are primarily interested in the relationships among macro elements (cf. pp. 66-70).

I think that in spite of this reasoning fitting the equation directly to the means of the replicates is problematic for several reasons.

Disregarding the number of replicates is an infringement of a fundamental maxim of regression analysis: "Use all the relevant data in estimating each constant" (Daniel, Wood 1980: 5). More specifically, failing to take the number of replicates into account shows that the following important fact has been disregarded: the higher the number

of replicate responses, the more reliable are their means; and the more reliable a mean is, the more weight it should be given when the parameters of the regression equation are estimated. (Hammerl 1991: 157f. partly acknowledges this fact by pooling x values with a small number of replicated responses.)

Furthermore, when group means are used as observations and the number of observations is not the same in every group, the disturbance is heteroskedastic and the ordinary least squares estimators are not efficient (cf. Kmenta 1971: 322ff.).

Finally, as has been pointed out in Section 2.7, the replicates provide the most reliable estimate of σ^2. In Köhler's and Hammerl's approach, when the regression equation is directly fitted to the means, the means are treated not as aggregates but as indiviudal observations which are all equally weighted with 1.0 (cf. also Draper and Smith's 1981: 278f. comments on "Regression Analysis with Summary Data" and Kmenta's 1971: 322ff. discussion of "estimation from grouped data").

If Köhler and Hammerl had taken the number of replicates into account, they would have obtained a different, and more accurate, regression equation. Since their samples are relatively large (Köhler uses part of the LIMAS corpus), the regression equation would have been far more reliable. The equation might then be checked with the F test for lack of fit, and, if the result is nonsignificant, with an F test for overall regression as described in Section 2.7 above. Note that in these tests, in contrast to the F test carried out by Köhler (1986), the total number of observations, N, and not only the number of the means of replicates is taken into account. The problem of large sample size might thus become crucial. Further note that Hammerl's use of R^2 is also problematic (cf. Sections 2.6.4, 3.2 and 3.5).

Finally, Köhler's interpretation of the results of his various F tests is also incorrect. The author uses the F test not only as a significance test, but explicitly characterizes the F statistic as a *measure* of goodness of fit (cf. in particular Köhler 1986: 99). As shown in Sections 2.4 and 3, this is not appropriate.

Equally inappropriate is the following interpretation of the F test by Hammerl (1991):

> The greater the calculated test statistic $F(f_1, f_2)$ is compared to the corresponding tabulated value for a given significance level α, the better is the fit of the theoretical values to the empirical data. (p. 154)

The inappropriateness of the F test as a measure of goodness of fit may be illustrated by Köhler's own data. Modelling the relationship between the predictor variable 'number of contexts in which a lexical entity is used' and the dependent variable 'frequency of the lexical entity', Köhler obtains the following result for the F test: $F_{1,122} = 332; p = 1.2 \cdot 10^{-11}$. Since the obtained F value is almost 50 times larger than the critical F value, Köhler concludes that the fit of the model is very good. However, as he points out himself, this conclusion seems to be not completely supported when one visually inspects the fitted curve.

This impression is confirmed if we transform the F value into the coefficient of determination R^2 according to equation (9). One obtains $R^2 = .73$, which indicates a fit far from perfect.

Another point of criticism concerns the illegitimate application of the F test when a method of estimation other than ordinary least squares is used. This criticism applies, for example, to the work of Altmann and Kind (1983) or Hammerl (1987). To predict the frequency of words of different levels of abstraction (Martin's 'law'; cf. Martin 1974), these authors use, among other methods, an estimation procedure based on the ratio of the two largest frequencies. They then employ the F test to evaluate the adequacy of their regression models. However, as has been shown in Section 2.3, the F test is only valid if the method of least squares (or maximum likelihood or best linear unbiased estimation) is used.[7]

[7] This criticism possibly also applies to the work of Köhler (1986), who characterizes the method of least squares as a "first approximation" compared to the procedure he uses himself (cf. p. 99). (Köhler is not very clear about his method of parameter estimation.)

3. The Coefficient of Determination

As has been shown in Section 2, the F test becomes significant when the sample size increases and is thus not an appropriate measure of the adequacy of regression models. The measure of goodness of fit and predictive power most often used for regression models is the coefficient of determination R^2. There are, however, various problems connected to this coefficient, some of which seem to be almost completely ignored in quantitative linguistics and even in many publications on regression analysis. A number of problems dealt with in the literature will now be briefly discussed.

3.1 Some Alternative Statistics

As has been pointed out by Kvålseth (1985), the following expressions for R^2 are found in the literature:

$$R_1^2 = 1 - \sum(y_i - \hat{y}_i)^2 / \sum(y_i - \bar{y})^2 \tag{31}$$

$$R_2^2 = \sum(\hat{y}_i - \bar{y})^2 / \sum(y_i - \bar{y})^2 \tag{32}$$

$$R_3^2 = \sum(\hat{y}_i - \bar{\hat{y}})^2 / \sum(y_i - \bar{y})^2 \tag{33}$$

$$R_4^2 = 1 - \sum(e_i - \bar{e})^2 / \sum(y_i - \bar{y})^2, \quad e_i = y_i - \hat{y}_i \tag{34}$$

$$R_5^2 = \text{squared multiple correlation coefficient between the regressand and the regressors} \tag{35}$$

$$R_6^2 = \text{squared correlation coefficient between } y \text{ and } \hat{y} \tag{36}$$

$$R_7^2 = 1 - \sum(y_i - \hat{y}_i)^2 / \sum y_i^2 \tag{37}$$

$$R_8^2 = \sum \hat{y}_i^2 / \sum y_i^2 \tag{38}$$

In the above expressions all summations are from $i = 1$ to $i = n$. $\bar{y}$ and $\bar{\hat{y}}$ denote the arithmetic mean of y_i and $\hat{y}_i$, respectively. Note that the definition of R_1^2 and R_2^2 is directly based on equations (8) and (7) above.

When we choose an appropriate R^2, we have to take into account (a) the type of model being fitted, (b) the fitting technique used, and (c)

the properties of R^2 considered to be desirable for a given purpose (cf. Kvålseth 1985: 280).

The choice is relatively unproblematic when a linear model as defined in equation (1) is used and when the parameters are estimated by means of the ordinary least squares method. In this case $R_1^2 = R_2^2 = \cdots = R_6^2$ (cf. Draper, Smith 1981, chaps. 1-2).

However, the value for these six statistics is generally different from the values for R_7^2 and R_8^2 which have been recommended by some authors for the case of linear no-intercept models, that is, for $\beta_0 = 0$ in equation (1) (cf., e.g., Montgomery, Peck, 1982: 38-43). In the case of no-intercept models, $R_7^2 = R_8^2$, whereas all other R^2 statistics generally yield different values (cf. Kvålseth 1985 and Judge, Griffiths, Hill, Lütkepohl, Lee 1985; the latter also discuss the calculation and interpretation of R^2 in models with correlated or heteroskedastic errors).

Furthermore, as has been pointed out by Helland (1987), R^2, which he defines as R_2^2, can be interpreted as an estimate of the population parameter only when the predictor variables are stochastic. However, even in the latter case, the sample coefficient of determination may be considerably above the highest approximate confidence limit for the population coefficient of determination as given by Helland (1987: 65).

Similarly, Assenmacher (1984: 118) argues that although the coefficient of determination is formally equivalent to the squared coefficient of correlation, it does not have the same inferential properties. Correlational analysis is based on the assumption that all observations are random samples from a multivariate normal distribution. This assumption is, however, as a rule not valid in regression analysis.

Another problem arises when there are replicated responses. In this case R^2 cannot attain 1 no matter how well the model fits the data. The problem of replicates will be discussed in some detail in Section 3.5.

Finally it should be noted that selection of a regression model from the data by procedures such as backward elimination, forward selection, stepwise selection or best subset regression leads to an inflation of R^2 and to biased partial F and t values (cf. Diehr, Hoflin 1974; Draper, Smith 1981: 294-312; Rencher, Pun 1980).

3.2 Some Problems with Nonlinear Models

According to Kvålseth (1985: 280) one of the most frequent mistakes with nonlinear models consists in comparing a linear with a nonlinear model that has been linearized, by using the same R^2 expression but different variables: the original y and the fitted $\hat{y}$ for the linear model and the transformed variables for the nonlinear model. The problem is that the R^2 value based on the transformed data provides a measure of fit for the linearized model rather than for the nonlinear model. As a consequence, a nonlinear model may be selected although the analysis of the residuals favours the linear model (cf. Kvålseth 1985 for a numerical example).

An interesting example is discussed in Heil (1987: 211-215). Heil shows for a specific exponential intercept model that if we linearize the model by means of a logarithmic transformation, subsequently estimate the parameters of the linearized model using the method of ordinary least squares, and then convert the linearized model back to the original model, the coefficient of determination is affected as follows: for the linearized model we obtain $R_1^2 = R_2^2 = 0.86$, and for the original model $R_1^2 = 0.79$ and $R_2^2 = 1.39$. We thus have different values for R_1^2 in the case of the linearized and the original model and furthermore a large difference between R_1^2 and R_2^2 for the nonlinear model. The latter fact shows that the decomposition of SST into SSR and SSE is not valid for the nonlinear model. This is due to the fact that the linearization by means of a logarithmic transformation results in an estimator for the intercept which does not possess the least squares properties (cf. Section 2.6 above).

In quantitative linguistics, this problem seems to be ignored, for example, by Hammerl (1991: 158ff.). Although Hammerl has linearized his nonlinear models by means of logarithmic transformations, he interprets R_1^2 in terms of the proportion of variance accounted for by the corresponding nonlinear model.

A further problem arises with R_5^2, which measures the strength of the correlation between the dependent variable and the predictor variables in a linearized model, but does not necessarily yield an appropriate measure of the fit of the original nonlinear model.

Finally, for intrinsically nonlinear models, the different R_i^2 generally yield different values. Furthermore, both R_2^2 and R_3^2 may be greater than 1 and R_6^2 may yield a high value even if the values for y and $\hat{y}$ deviate substantially (cf. Kvålseth 1985: 280).

3.3 The Adjusted Coefficient of Determination

Sometimes it may be preferable to take into account the degrees of freedom of the model being fitted. This may be advisable, for example when the goodness of fit of models differing in the number of degress of freedom is compared. In this case, the use of R^2 may be misleading since increasing the number of predictors will usually lead to an increase of R^2 even when the true values of the new regression coefficients are zero.

Furthermore, when the number of predictor variables is large compared to the sample size, R^2 tends to overestimate the relationship in the population. As Helland (1987) has pointed out, in contrast to the unadjusted R^2, the adjusted R^2 (Helland acutally refers to R_2^2) will always be inside the approximate confidence limits for the population coefficient of determination as established by Helland (1987: 65). The adjusted R^2 is closely connected to Mallows C_p statistic, which is sometimes used as an alternative to R^2, for example in best subset regression (cf., e.g., Draper, Smith 1981: 299ff.).

To adjust R_1^2 for the appropriate degrees of freedom, we may divide $\sum(y_i - \hat{y}_i)^2$ and $\sum(y_i - \bar{y})^2$ by their corresponding degrees of freedom. (I consider only R_1^2 because there is evidence that it should be preferred to the coefficients R_2^2 to R_8^2 (cf. Section 3.6 below).) This results in the following adjusted coefficient of determination:

$$\begin{aligned} \overline{R}^2 &= 1 - a\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= 1 - a(1 - R_1^2) \end{aligned} \tag{39}$$

with the adjustment factor $a = (n-1)/(n-J-1)$ for the linear intercept model and $a = n/(n-J)$ for a no-intercept model (for a discussion of various adjustments cf. Carter 1979).

$\overline{R}^2$ is always less than or equal to R_1^2. Asymptotically the two measures are equal. Unlike R_1^2, which is always positive (unless the model is grossly misspecified), $\overline{R}^2$ may have negative values.

3.4 The Interpretation of R^2

In the case of a general linear regression model and ordinary least squares estimation, $SST = SSR + SSE$ and R^2 is to be interpreted as a measure of the proportion of variance of the regressand attributable to sample regression.

Note that if we use a different method of estimation, as is sometimes the case in quantitative linguistics, the decomposition of SST into SSR and SSE is not possible and R^2 can no longer be interpreted as a measure of the proportion of variance of the regressand attributable to sample regression. Instead, it must be interpeted purely as a measure of the goodness of fit (cf. Kmenta 1971: 234). However, as Heil (1987: 215ff.) has demonstrated, R^2 can be highly misleading in this case. (In Heil's example, which is admittedly extreme, $R_1^2 = -3$ while $R_2^2 = 1$!)

The problem of estimation is not taken into account, for example, by Hammerl (1990: 12f.) when dealing with Arapov's model of the relationship between word length and rank. Although Hammerl uses an iterative procedure for parameter estimation, he interprets the coefficient of determination as a measure of the amount of variation explained.

Similarly, if the linearization of a nonlinear model involves the transformation of the parameters to be estimated, R^2 adequately measures only the proportion of variance explained by the linearized model, and not that accounted for by the nonlinear model (cf. Section 3.2 above).

If, however, the linearization leads to a no-intercept model, R^2 is not an adequate measure even for the linearized model, because for linear no-intercept models the decomposition of SST into SSR and SSE is not valid. The latter fact is shown, for example, by Heil (1987: 217ff.) and Kvålseth (1985). (Heil discusses an example of a linear no-intercept model where $R_1^2 = 0.89$ and $R_2^2 = 1.6$.)

But also in the case of linear models and ordinary least squares estimation, the interpretation of R^2 involves a number of problems. I will first consider the case of only one predictor variable. Note that most

of what will be said also applies when several predictor variables are involved. Subsequently, special problems arising in the case of multiple predictors will be discussed.

3.4.1 Simple Linear Regression

Suppose we find a very low value for R^2 when fitting a regression line to a given sample. According to Kmenta (1971) there are three possible explanations for the low value of R^2. First, one might argue that x is a poor explanatory variable in that it leaves variation in y unaffected. Another possible explanation is that x is the relevant explanatory variable, but that its influence on y is weak compared to the influence of random error. A third possible explanation is that the regression equation is misspecified. According to Kmenta (1971) this is the conclusion that is often reached in practice: the value of R^2 is taken as an indication of the "correctness" of the specification of the model. Whereas the first two interpretations might be checked with the help of inferential statistics, the latter one involves a purely operational criterion that has no foundation in statistical inference (cf. Kmenta 1971: 234).

In this context it should be noted that strictly speaking an inferential interpretation of R^2 presupposes that *all* variables involved are stochastic (and not only the regressand) and that furthermore *all* observations are random samples from a multivariate normal distribution (cf. Assenmacher 1984: 118 and Helland 1987). For the regressand this assumption might be often warranted, at least asymptotically. It is, however, only seldom justified to assume that all regressors follow a multivariate normal distribution. This problem is compounded in quantitative linguistics where most variables follow distributions which are quite different from a normal distribution.

But even if we treat R^2 as a purely descriptive measure, its interpretation is by no means straightforward. For example, if the assumption of a multivariate normal distribution of all variables is not fulfilled, the following problem may arise: if the distributions of the variables are quite dissimilar from each other, for example, if one distribution is heavily skewed to the right and the other to the left, then the maximum that $|r|$ can attain may be considerably less than 1 (cf., e.g., Carroll 1961; Diehl, Kohr 1991, chap. 10). As a consequence, R^2 may

be relatively small although the relationship between the regressand and the set of regressors might be quite strong.

Similarly, if the assumption of homoskedasticity is not fulfilled, the interpretation of R^2 is highly problematic. If the regression is heteroskedastic, the amount of variance explained is not constant for the whole range of the predictor variable: for one part of the responses, the prediction may be good, for another part quite poor. However, R^2 measures only the *average* amount of variation explained. Therefore, the value obtained may be quite misleading for part of the observations (cf. Heil 1987: 208 for a similar argument).

A problem, which is closely related to Kmenta's second point above, concerns the dependency of R^2 on how reliably the variables are measured. Suppose we want to compare the adequacy of one and the same model for different populations (e.g., different languages). If it is possible to measure the variables more reliably in one of the data sets investigated for this purpose, the error variance in this sample is reduced and R^2 is increased. This does not mean, however, that the influence of the independent variable(s) on the dependent variable is also necessarily stronger in this particular sample compared to the other samples (cf. also Urban 1982: 54).

Another problem has been pointed out by Barrett (1974) who argues that R^2 can be viewed as a measure of *both* the goodness of fit in the sense of how close the data fit the regression surface, and the steepness of the regression surface. If the goodness of fit about the regression surface is taken to be fixed, then with a steeper surface $\sum_{i=1}^{n}(y_i - \bar{y})^2$ will become larger and thus increase R^2. Hence, predictions based on a regression equation with a large R^2 and a steep regression surface might not be more precise and could even be less precise than the predictions based on an equation with a smaller R^2 and with a less steep surface (cf. Barrett 1974, p. 19).

Barrett (1974) illustrates this fact by rotating the regression surface about $\bar{x}$ and $\bar{y}$ from 0 to 90 degrees while holding constant the vertical distances of the data points to the surface (i.e., $\sum_{i=1}^{n}(y_i - \hat{y})^2$. This results in a rapid increase of R^2, whereas the predictive precision remains constant (except for a slope of 0 degrees).

An illustrative example of some problems involved in the interpretation of R^2 in the case of one predictor variable is given by Anscombe (1973). Anscombe presents four different data sets, each consisting of eleven (x, y) pairs and each yielding identical statistics (regression coefficients, residual sums of squares, R^2, etc.) when a straight line is fitted. For each data set, $R^2 = 0.67$. However, it is only for the first data set that the assumptions of the linear model are fulfilled and R^2 adequately describes the data. The second data set shows a nonlinear relationship between x and y that would not have been detected by relying solely on the large value of R^2. The remaining two data sets are, though in very different ways, strongly affected by an outlier, and the corresponding regression line (being the same for both data sets) does not describe the bulk of the data very well in spite of the large value of R^2. This is particularly true for the fourth data set (cf. also Urban 1982: 52-57).

3.4.2 Multiple Linear Regression

In the case of multiple linear regression additional problems arise. If the predictor variables are uncorrelated with each other

$$R^2 = r^2_{yx_1} + r^2_{yx_2} + \cdots + r^2_{yx_J}. \tag{40}$$

However, when the predictor variables are intercorrelated, R^2 can be both less and greater than $r^2_{yx_1} + r^2_{yx_2} + \cdots + r^2_{yx_J}$. When the predictive power of a variable x_j or set of variables $\{x_j\}$ is increased by the inclusion of another predictor variable which is correlated with the variable x_j or the set of variables $\{x_j\}$, the included variable is called a suppressor variable. It is a suppressor for those variables whose regression weights are increased (cf. Conger 1974: 36f.; Keeves 1988; Tzelgov, Stern 1978). We are dealing with a classical suppression effect when in the case of two predictor variables one predictor has a correlation of zero with the criterion but has a significant regression coefficient when the criterion is regressed on both predictor variables (cf. Hamilton 1987; Keeves 1988).

In the case of several predictor variables it is thus as a rule very difficult to determine the contribution of each predictor variable to R^2. As has already been pointed out in the discussion of multicollinearity in Section 2.4.5, simply looking at the coefficients of correlation or at the plots of the x_j against y may be highly misleading.

It has been argued in the pertinent literature that suppressor variables should be used to increase the predictive power of a regression model. This appears justified, whenever optimization of prediction is the main objective. In quantitative linguistics, however, the primary and ultimate aim is to develop explanatory models. In this situation it seems to be more adequate to include a suppressor variable in a regression equation only if the inclusion is strongly supported on theoretical grounds.

Some of the problems just outlined are also discussed by Korn and Simon (1991). The authors distinguish between two different properties of a regression model: (1) the consistency of the model with the data – historically referred to as "goodness of fit", and (2) the ability of the predictor variables to distinguish different outcomes. The second property is referred to by Korn and Simon (1991) as "explained risk" and defined as the proportional decrease in risk obtained by using the predictor variables. The explained risk is a population quantity which depends only on the model and the distribution of the predictor variables but not on the observed values of y.

Korn and Simon (1991) discuss various estimators of explained risk. One possible estimator is the coefficient of explained residual variation R^2. If the model is misspecified, the explained residual variation is a large-sample underestimate of the risk explained by the correctly specified model.

3.4.3 Goodness of Fit vs. Explained Risk

It may happen that the fit of a model is perfect, while the explained risk is low (cf. Korn, Simon 1991: 201 for examples). This may be the case when there are replicated responses (cf. Section 3.5 below). However, if we view R^2 as a measure of explained risk, it is not embarrassing that R^2 is less than 1 even if the fit of the model is perfect: the value obtained simply means that the predictor variables have not explained all the risk (cf. Korn, Simon 1991: 205).

In quantitative linguistics, the fact that R^2 measures both goodness of fit and explained risk seems to be overlooked, for example, by Hammerl (1990), who seems to interpret low values of R^2 primarily as an indication that predictor variables which might contribute substantially to the explanation of the variance of the dependent variable have not been

included in the model (cf. pp. 9f.). In the light of what has been said above, this interpretion appears to be quite simplistic.

In view of the situation described it seems to me to be also quite problematic to compare R^2 with preestablished conventional boundaries as has been done by Hammerl (1991: 160), who considers the fit of his models to be satisfactory if $R^2 \geq 0.9$.

3.5 Replicated Responses

It has been shown by Draper (1984) that R^2 (defined as R_1^2) can be made small simply by increasing the number of replicate values of y at existing values of x, although the fit of the model is at the same time improved. Moreover, the achievable upper bound of R^2 is below 1.0 when there are replicated responses. A perfect fit thus does not correspond to $R^2 = 1$. For these reasons, R^2 is misleading as a measure of fit in the presence of replicated responses (cf. also Healy 1984).

The latter fact can easily be proved on the basis of the results presented in Section 2.7 (for the following cf. Chang, Afifi 1987). It has been shown that $SST = SSR + SSE$ and $SSE = SSW + SSL$. Hence

$$SST = SSW + SSL + SSR. \tag{41}$$

From ANOVA we know that SST can also be partitioned as

$$SST = SSW + SSB, \tag{42}$$

where $SSB = \sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2$ is the sum of squares between means of replicates. It immediately follows that

$$SSB = SSL + SSR. \tag{43}$$

Equation (43) indicates that $SSR \leq SSB$. The most that any regression equation can explain in the case of replicated responses is thus the variation among the means of the replicates.

Since SSW is not affected by the regression equation fitted to a given set of data,

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{SSL + SSW}{SST} \tag{44}$$

is less than 1.0 whenever $SSW \neq 0$ (cf. also Draper, Smith 1981: 547f.).

Equation (43) implies that $R^2 = SSR/SST = (SSB - SSL)/SST$. Thus $R^2 \leq SSB/SST$. SSB/SST represents an achievable upper bound on R^2; it is also known as the "correlation ratio" or "eta square" (cf. Chang, Afifi 1987: 196).

Equation (43) suggests that in the case of replicated responses, R^2 might be modified as follows:

$$R^2_M = \frac{SSR}{SSB} = 1 - \frac{SSL}{SSB}. \tag{45}$$

The modified coefficient of determination R^2_M is 1.0 when $SSL = 0$, that is, when the regression equation fits the means of the replicates perfectly. It is 0 when $SSL = SSB$, which according to equation (43) implies that $SSR = 0$. When there are no replicates, $R^2_M = R^2$.

If we let $R^2_U = SSB/SST$ denote the achievable upper bound of R^2, then R^2_M can also be expressed as $R^2_M = R^2/R^2_U$. Thus R^2_M might also be interpreted as measuring how well a given regression equation fits the data compared to the best possible equation.

Chang and Afifi (1987: 196) point out that R^2_M is more effective than R^2 when the purpose is to determine whether including some higher-order terms of the predictor variables (i.e., cross-products or squares) would improve the fit. However, R^2_M is not useful for *selecting* predictor variables, since removing or adding predictor variables results in distinct explanatory vectors, and hence in different compositions of the replicates and different values of K, SSB, SSW, and R^2_U.

R^2_M might also be adjusted for the appropriate degrees of freedom (cf. Section 3.3 above). Let df_T, df_W, df_B, and df_L denote the number of degrees of freedom of SST, SSW, SSB, and SSL. According to Table 1 in Section 2.7, $df_L = K - J - 1$. Furthermore, since $SSB = SST - SSW$, $df_B = df_T - df_W = N - 1 - (N - K) = K - 1$. Hence R^2_M might be modified as follows (cf. Chang, Afifi 1987: 196):

$$\overline{R}^2_M = 1 - \frac{SSL/(K - J - 1)}{SSB/(K - 1)}. \tag{46}$$

Other possible goodness-of-fit measures for the special case of replicated responses are discussed, for example, in Draper (1984) or Healy

(1984). A comparison of different measures is presented in Chang and Afifi (1987).

In Section 2.8 it was pointed out that Hammerl (1991) fitted his regression models directly to the means of the replicate responses thus disregarding the number of replicates, and then after linearization of his models used the ordinary coefficient of determination as defined in equation (31) to evaluate the adequacy of his models. I have argued that disregarding the number of replicates is an infringement of the maxim that all the relevant data should be used in estimating each parameter. In accordance with this argument I consider it more appropriate to take the number of replicates into account both in parameter estimation and in model evaluation. For the latter purpose, if there are replicate responses, one might calculate both the ordinary coefficient of determination R^2 and the modified coefficient R^2_M (both possibly adjusted for the corresponding degrees of freedom). R^2 could then be interpreted as a measure of the predictive power of the model with regard to the individual responses (Hammerl's micro elements), whereas R^2_M could be interpreted as a measure of lack of fit as well as of predictive power with regard to the means of the replicates (Hammerl's macro elements).

3.6 Possible Guidelines

In view of the quite confusing situation outlined so far, one might ask whether any guidelines may be formulated as to the use of R^2 in regression analysis.

There are various, often contradictory, suggestions in the literature. After a thorough discussion of potential pitfalls in using the different R^2 statistics, Kvålseth (1985) suggests that R^2_1 be consistently used both for intercept and no-intercept models as well as for (intrinsically) linear and intrinsically nonlinear models. R^2_1 is chosen by Kvålseth because this coefficient has the largest number of desirable properties as compared to the other coefficients. By recommending the use of only one coefficient Kvålseth attempts to provide for comparability when different models are fitted to the same set of data.

Kvålseth also notes, however, some restrictions with regard to R^2_1: it may be negative when the model being fitted is grossly misspecified

(e.g., when a linear no-intercept model is fitted to data for which β_0 is substantially different from 0) or when a fitting technique such as a robust regression method is used for data with gross outliers (for the identification of outliers and for robust model-fitting methods cf., e.g., Hawkins 1980 and Huber 1981). The first case presents no problem (a negative value simply indicating a complete lack of fit). The second case, however, is problematic since even a single outlier may have a dramatic effect on R_1^2, thus indicating poor fit when the fit is actually very good.

In the case of robust regression, Kvålseth (1985: 283) suggests modifying R_1^2 by replacing the arithmetic means by sample medians. This results in a coefficient of determination R_9^2 which is robust against outliers. This coefficient can also be adjusted for the number of degrees of freedom (cf. Kvålseth 1985: 284).

In sum, Kvålseth's (1985) recommendation reads as follows: Use R_1^2 or $\overline{R}^2$ consistently for any type of model- and curve-fitting technique except for the case in which a robust model-fitting technique is applied to highly contaminated data. In this latter case use the robust determination coefficient R_9^2. Supplement the calculation and interpretation of R^2 by a detailed analysis of the residuals.

Although analysis of residuals is recommended throughout the pertinent statistical literature (comprehensive treatments can be found in Belshey, Kuh, Welsch 1980; Cook, Weisberg 1982; or Hawkins 1980), and residual scatterplots are provided by all major statistical programs (e.g., SPSS, BMDP, SAS, SYSTAT), analysis of residuals is quite an exception in quantitative linguistics. I am sure that at least in some cases a thorough analysis of the residuals would have led to a different evaluation of the models fitted.

If there are replicated responses I would suggest that the modified coefficient of determination R_M^2 (or $\overline{R}_M^2$) should be calculated in addition to R^2. This would allow us to measure both the goodness of fit of a given regression equation and the predictive power of the equation with regard to the means of the replicated responses.

4. Testing the Stability of Parameters Over the Sample Space

It is important for the validity of a regression model that the parameters are stable over the whole range of the sample. With longitudinal data, that is, with data collected over an extended period of time, the stability of the estimated parameters can be tested by fitting the model to shorter time spans, and comparing the pattern of consecutive estimates of the regression coefficients. If the coefficients are not stable over time, the use of a regression model estimated from *all* the data for prediction would be problematic (cf. Draper, Smith 1981: 419).

With cross-sectional data, that is, with data collected at a single point in time, the method of cross-validation can be used to check the stability of the parameter estimates (cf. for the following Snee 1977; Stelzl 1982: 124-138; Stone 1974; Wainer 1978). Although many textbooks on regression analysis highly recommend the use of this method, it seems to be almost completely ignored in quantitative linguistics.

Cross-validation involves using two different, though comparable, samples from the same population. These are often obtained by splitting a given sample.[8] The first sample, the so-called estimation data, is used for parameter estimation. The regression equation obtained is then used to predict the data of the other sample, called the prediction data. The correlation between the predicted and observed values provides a measure of the predictive power of the regression equation obtained.

Cross-validation is recommended because (multiple) regression capitalizes on sample specific variation and chance differences. Whenever a regression model is fitted to a sample, for example by means of the method of least squares, the model is optimal for that sample. However, optimum fit in one sample does not mean that the fit is also optimal in another sample from the same population. Usually the correlations between predictors and criterion as well as the correlations among predictors will fluctuate from sample to sample. Furthermore, in procedures such as stepwise or setwise regression, decisions about which variables are included or ommitted depend on chance differences

[8] Methods for data splitting are discussed, e.g., in Snee (1977).

within a single sample. As a consequence, the equation derived from a sample may overfit the data and poorly generalize to the population. Yet in order for a model to be generally useful, it should have reasonably good extrapolation properties (cf. in particular Snee 1977 and Wainer 1978). For these reasons, correlating the predicted and the observed values as described above provides a more realistic measure than, for example, the coefficient of determination R^2 calculated on the basis of a single sample.

One potential drawback of data splitting is that the variances of the coefficients calculated from a subsample will be larger than those from the total sample. However, since in quantitative linguistics the sample size is often very large, the variances of the estimates will be small enough, even if only half of the sample is used for estimation. I would suggest that cross-validation be used much more often in quantitative linguistics.

5. Conclusion

It has been shown in this article that the use of the F test (as well as other significance tests) in regression analysis involves a number of problems and that as a consequence an evaluation of the adequacy of a regression model solely by means of the F test is inappropriate. It has been pointed out that the F test (and also other tests) may be significant solely because the sample size is large. Furthermore, it has been demonstrated that the F test is based on assumptions which are often not fulfilled in quantitative linguistics, and that it is in general not valid in the case of (intrinsically) nonlinear models as well as with replicate responses (at least in its ordinary form).

It has been suggested that the coefficient of determination R^2 should be used as a measure of the consistency of the model with the data, and of the ability of the predictor variables to distinguish different outcomes. It has been demonstrated, however, that the use and interpretation of R^2 involves a number of potential pitfalls. For example, it has been shown that the value of R^2 obtained for a specific sample depends on various factors such as the kind of model fitted, the existence of replicated responses, the existence of outliers, the degree of multicollinearity, the steepness of the regression surface, and the number of predictor variables as compared to the sample size.

In sum, I would like to suggest that the evaluation of a regression model should be based on as much information as can be made available. The evaluation may include carrying out an overall F test, and individual F tests or t tests for each parameter, construction of confidence regions, computation of coefficients of multiple and (semi-)partial determination, analysis of residuals, checking of stability of the parameters over the sample space, comparison of results with theoretical models and simulated data, and last, but not least, sound theoretical reasoning. For, as Draper and Smith (1981: 422) put it: "The use of multiple regression techniques is a powerful tool only if it is applied with intelligence and caution." I think that the above suggestions can contribute to a further improvement of the construction of regression models in quantitative linguistics as well as in other empirical sciences.

References

Altmann, G. (1983), "Das Piotrowski-Gesetz und seine Verallgemeinerungen". In: Best, K.-H., Kohlhase, J. (eds.), *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte.* Göttingen: Herodot, 59-90.

Altmann, G. (1992), "Phoneme counts. Comments on Pääkkönen's article". (Ms.).

Altmann, G., von Buttlar, H., Rott, W., Strauß, U. (1983), "A law of change in language". In: Brainerd, B. (ed.), *Historical linguistics.* Bochum: Brockmeyer, 104-115.

Altmann, G., Kind, B. (1983), "Ein semantisches Gesetz". In: R. Köhler, J. Boy (eds.), *Glottometrika 5.* Bochum: Brockmeyer, 1-12.

Amemiya, T. (1983), "Non-linear regression models". In: Griliches, Z., Intriligator, M.D. (eds.), *Handbook of econometrics. Vol. 1.* Amsterdam: North Holland, 334-389.

Anscombe, F.J. (1973), "Graphs in statistical analysis". *The American Statistician* 27, 17-21.

Assenmacher, W. (1984), *Einführung in die Ökonometrie.* München: Oldenbourg.

Barrett, J.P. (1974), "The coefficient of determination – some limitations". *The American Statistician* 28, 19-20.

Belsley, D.A., Kuh, E., Welsch, R.E. (1980), *Regression diagnostics.* New York: Wiley.

Binkley, J.K., Abbott, P.C. (1987), "The fixed X assumption in econometrics: Can the textbooks be trusted?" *The American Statistician* 41, 206-214.

Box, G.E., Draper, N.R. (1987), *Empirical model-building and response surfaces.* New York: Wiley.

Carroll, J.B. (1961), "The nature of the data, or how to choose a correlation coefficient". *Psychometrika* 26, 347-372.

Carter, D. (1979), "Comparison of different shrinkage formulas in estimating population multiple correlation coefficients". *Educational and Psychological Measurement* 39, 261-266.

Chang, P., Afifi, A. (1987), "Goodness-of-fit statistics for general linear regression equations in the presence of replicated responses". *The American Statistician* 41, 195-199.

Cohen, J. (1977), *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

Conger, A.J. (1974), "A revised definition of suppressor variables: A guide to their identification and interpretation". *Educational and Psychological Measurement* 34, 35-46.

Cook, R.D., Weisberg, S. (1982), *Residuals and influence in regression.* London: Chapman and Hall.

Daniel, C., Wood, F.S. (1980), *Fitting equations to data.* New York: Wiley.

Diehl, J.M., Kohr, H.U. (1991), *Deskriptive Statistik* (9th ed.). Eschborn: Klotz.

Diehr, G., Hoflin, D. (1974), "Approximating the distribution of sample R^2 in best subset regression". *Technometrics* 16, 317-320.

Dillon, W.R., Goldstein, M. (1984), *Multivariate analysis: Methods and applications.* New York: Wiley.

Draper, N.R. (1984), "The Box-Wetz Criterion versus R^2". *Journal of the Royal Statistical Society A* 147, 100-103.

Draper, N., Smith, H. (1981), *Applied regression analysis* (2nd ed.). New York: Wiley.

Fahrmeier, L., Kaufmann, H., Kredler, C. (1984), "Regressionsanalyse". In: Fahrmeier, L., Hamerle, A. (eds.), *Multivariate statistische Verfahren.* Berlin: de Gruyter, 83-153.

Grotjahn, R., Altmann, G. (in press), "Modelling the distribution of word length: Some methodological problems". Paper presented at the First Quantitative Linguistics Conference (QALICO), Trier, September 23-27, 1991. (To appear in the QALICO Proceeding. Amsterdam: Elsevier 1992).

Hamilton, D. (1987), "Sometimes $R^2 > r^2_{yx_1} + r^2_{yx_2}$. Correlated variables are not always redundant". *The American Statistician* 41, 129-132.

Hammerl, R. (1987), "Untersuchungen zur mathematischen Beschreibung des Martingesetzes der Abstraktionsebenen". In: Fickermann, I. (ed.), *Glottometrika 8.* Bochum: Brockmeyer, 113-129.

Hammerl, R. (1990), "Länge-Frequenz, Länge-Rangnummer: Überprüfung von zwei lexikalischen Modellen". In: Hammerl, R. (ed.), *Glottometrika 12.* Bochum: Brockmeyer, 1-24.

Hammerl, R. (1991), *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Trier: Wissenschaftlicher Verlag Trier.

Hawkins, D. (1980), *Identification of outliers.* London: Chapman and Hall.

Healy, M.J.R. (1984), "The use of R^2 as a measure of goodness of fit". *Journal of the Royal Statistical Society A* 147, 608-609.

Heil, J. (1987), *Einführung in die Ökonometrie* (2nd ed.). München: Oldenbourg.

Helland, I.S. (1987), "On the interpretation and use of R^2 in regression analysis". *Biometrics* 43, 61-69.

Huber, P.J. (1981), *Robust statistics.* New York: Wiley.

Judge, G.G., Griffiths, W., Hill, R., Lütkepohl, H., Lee, T.-C. (1985), *The theory and practice of econometrics* (2nd ed.). New York: Wiley.

Judge, G.G., Hill, R., Griffiths, W., Lütkepohl, H., Lee, T.-C. (1982), *Introduction to the theory and practice of econometrics*. New York: Wiley.

Keeves, J.P. (1988), "Suppressor variables". In: Keeves, J.P. (ed.), *Educational research, methodology, and measurement. An international handbook*. Oxford: Pergamon, 774-775.

Kerlinger, F.N., Pedhazur, E.J. (1973), *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston.

Kinal, T., Lahiri, K. (1983), "Specification error analysis with stochastic regressors". *Econometrica* 51, 1209-1219.

King, M.L (1987), "Testing for autocorrelation in linear regression models: A survey". In: King, M.L., Giles, D.E.A. (eds.), *Specification analysis in the linear model*. London: Rutledge and Kegan Paul, 19-73.

Kmenta, J. (1971), *Elements of econometrics*. New York: Macmillan.

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

Korn, E.L., Simon, R. (1991), "Explained residual variation, explained risk, and goodness of fit". *The American Statistician* 45, 201-206.

Krämer, W., Kiviet, J., Breitung, J. (1990), "The null distribution of the F-test in the linear regression model with autocorrelated disturbances". *Statistica* 50, 503-509.

Krämer, W., Zeisel, H. (1990), "Finite sample power of linear regression autocorrelation tests". *Journal of Econometrics* 43, 363-372.

Kvålseth, T.O. (1985), "Cautionary note about R^2". *The American Statistician* 39, 279-285.

Marquardt, D.W., Snee, R.D. (1975), "Ridge regression in practice". *The American Statistician* 29, 3-20.

Martin, R. (1974), "Syntaxe de la définition lexicographique: étude quantitative des définissants dans le 'Dictionnaire fondamental de la langue française'". In: David, J., Martin, R. (eds), *Statistique et linguistique*. Paris: Klincksieck, 61-71.

McCullagh, P., Nelder, J.A. (1983), *Generalized linear models*. London: Chapman and Hall.

Montgomery, D.C., Peck, E.A. (1982), *Introduction to linear regression analysis*. New York: Wiley.

Pindyck, R.S., Rubinfeld, D.L. (1981), *Econometric models and economic forecasts*. New York: McGraw-Hill.

Rao, P., Miller, R.L. (1971), *Applied econometrics*. Belmont, CA: Wadsworth.

Rayner, J.C.W., Best, D.J. (1989), *Smooth tests of goodness of fit*. New York/Oxford: Oxford University Press.

Rencher, A.C., Pun, F.C. (1980), "Inflation of R^2 in best subset regression". *Technometrics* 22, 49-53.

Rützel, E. (1976), "Zur Ausgleichsrechnung: Die Unbrauchbarkeit von Linearisierungsmethoden beim Anpassen von Potenz- und Exponentialfunktionen". *Archiv für Psychologie* 128, 316-322.

Sampson, A.R. (1974), "A tale of two regressions". *Journal of the American Statistical Association* 69, 682-689.

Schneeweiß, H. (1974), *Ökonometrie* (2nd ed.). Würzburg: Physica.

Schönfeld, P. (1969), *Methoden der Ökonometrie. Bd. 1: Lineare Regressionsmodelle*. Berlin: Vahlen.

Snee, R.D. (1977), "Validation of regression models: Methods and examples". *Technometrics* 19, 415-428.

Stelzl, I. (1982), *Fehler und Fallen der Statistik – für Psychologen, Pädagogen, Sozialwissenschaftler*. Bern: Huber.

Stone, M. (1974), "Cross-validating choice and assessment of statistical predictions (with discussion)". *Journal of the Royal Statistical Society B* 36, 111-147.

Tabachnick, B.G., Fidell, L.S. (1989), *Using multivariate statistics* (2nd ed.). Cambridge: Harper & Row.

Tzelgov, J., Stern, I. (1978), "Relationships betwen variables in three variable linear regression and the concept of suppressor". *Educational and Psychological Measurement* 38, 325-335.

Urban, D. (1982), *Regressionstheorie und Regressionstechnik.* Stuttgart: Teubner.

Wainer, H. (1978), "On the sensitivity of regression and regressors". *Psychological Bulletin* 85, 267-273.

Wetz, J.M. (1964), *Criteria for judging adequacy of estimation by an approximating response function.* Ph.D. thesis, University of Wisconsin.

Willan, A.W., Watts, D.G. (1978), "Meaningful multicollinearity measures". *Technometrics* 20, 407-412.

Zörnig, P., Köhler, R., Brinkmöller, R. (1990), "Differential equation models for the oscillation of the word length as a function of the frequency". In: Hammerl, R. (ed.), *Glottometrika 12.* Bochum: Brockmeyer, 25-40.

Rieger, Burghard (ed.)
Glottometrika 13
Bochum: Brockmeyer 1992, 173-185

From character to word monkey, and some interesting applications [1]

Jacques B.M. Guy

Overview

An algorithm for computing the character or word-entropy of texts to any order has been summarily described in Guy 1990:125-130. An implementation of it for the computation of word entropy is described here in some greater detail. The data structure used allows not only the generation of text, but also text retrieval, the production of concordances, and appears to provide a useful means of research into text compression and the segmentation of continuous text into its constituent morphs.

The original algorithm was discovered while attempting to optimize for speed a class of text-generating algorithms described by Shannon[2]. As I was implementing it in Turbo Pascal, it became evident that it could be made to function also as an on-screen concordance and a simple text-retrieval program with little

---

[1] The permission of the Executive General Manager, Telecom (Australia) Research Laboratories, to publish this material is gratefully acknowledged.

[2] Dewdney (1989) reported a third-order word monkey operating on the same principle ("MARK V. SHANEY", a computer program created by B. Ellis on an idea of D.P. Mitchell of the AT&T Bell Laboratories):
"As MARK V. SHANEY scans a text, it builds a frequency table for all words that follow all the word pairs in the text. The program then proceeds to babble probabilistically on the basis of the word frequencies." A.K. Dewdney (1989:97)
As Dewdney does not elaborate on how the frequency table is built and stored one can only conjecture that it is held in memory perhaps as a sparse matrix (probably a B-tree), possibly with some recourse to disk storage.

additional effort. It is as I was writing the necessary code that further expansions to the algorithm occurred to me. Consequently, it is probably most clearly presented through its development stages, starting from text generation.

Text Generation

Say we wish to build a nth-order word monkey to ape this text: "He does his best, he does, he does his worst, he does".

Let n=3 (for instance).

Append n-1 null words (*) to the text, and number its characters:

```
"He  does his best,    he  does, he  does  his worst,  he  does    * * "
       1           2           3          4          5
12345678901234567890123456789012345678901234567890123456789012345678901234567
```

Draw up a list of all twelve three-word sequences contained in the text, referenced by the position of their first letter[3]:

Pos.	Pointing to:		
1	he	does	his
4	does	his	best
9	his	best	he
13	best	he	does
19	he	does	he
22	does	he	does
28	he	does	his
31	does	his	worst
36	his	worst	he
40	worst	he	does
47	he	does	*
50	does	*	*

[3] It ought to go without saying only that the input text and the pointers to the initial characters of each of its words are stored. The text pointed to is only given in the explanatory tables for the readers' convenience.

Sort them into alphabetical order:

Pos.	Pointing to:		
13	best	he	does
22	does	he	does
4	does	his	best
31	does	his	worst
50	does	*	*
19	he	does	he
1	he	does	his
28	he	does	his
47	he	does	*
9	his	best	he
36	his	worst	he
40	worst	he	does

Apply Shannon's algorithm restated for word approximations:

> "Let one select a sequence of n words at random from a text file. This string is output and stored in a variable called the 'context'. The file is then accessed at random and read until a word sequence identical to the context is encountered. The succeeding word is then output; it is appended to the context while the first word of the context is deleted. The file is again accessed at random, searched for this context and the succeeding word is output, etc."

In order to optimize this algorithm for speed we use precisely the same method as in the character-level text generator: record the frequency of occurrence of each context in the table at its place of first occurrence, and record a negative offset to the first occurrence in the other cells. Where a context contains the null word (*) enter a frequency of 0. Thus:

Pos.	Frequency or offset	Pointing to: Context		Word
13	1	best	he	does
22	1	does	he	does
4	2	does	his	best
31	-1	does	his	worst
50	0	does	*	*
19	4	he	does	he
1	-1	he	does	his
28	-2	he	does	his
47	-3	he	does	*
9	1	his	best	he
36	1	his	worst	he
40	1	worst	he	does

A binary search locates the desired context. If its frequency is negative, then it is the offset to the first occurrence of that particular context, where the true frequency is found, and from which its last occurrence is readily computed.

Computing the Word Entropy

The third-order word entropy of the text is computed from the data above in precisely the same manner as the character entropy of the original algorithm, the contribution of a word to the total entropy being

$$-\text{Log2}(\text{fqW}) * \text{fqW} * \text{fqC}$$

where fqC is the relative frequency of a context, and fqW the relative frequency of that word in that particular context. So that we have:

Pos.	Frequency or offset	Context		Word	fqC	fqQ	-Log2 (fqW) * fqW * FqC
13	1	best	he	does	1/11	1/1	0.0000
22	1	does	he	does	1/11	1/1	0.0000
4	2	does	his	best	2/11	1/2	0.0909
31	-1	does	his	worst		1/2	0.0909
50	0	does	*	*	n/a		
19	4	he	does	he	4/11	1/4	0.1818
1	-1	he	does	his		2/4	0.1818
28	-2	he	does	his			
47	-3	he	does	*		1/4	0.1818
9	1	his	best	he	1/11	1/1	0.0000
36	1	his	worst	he	1/11	1/1	0.0000
47	1	worst	he	does	1/11	1/1	0.0000
Sum of frequencies: 11						Third-order word entropy:	0.7273

As in the case of the character-level algorithm, in order to calculate lesser-order values of the entropy one only needs to adjust the size of the context and to recompute context frequencies. Thus for instance for the second-order word entropy, the context is one word long, and we have:

Pos.	Frequency or offset	Context		Word	fqC	fqQ	-Log2 (fqW) * fqW * FqC
13	1	best	he	does	1/12	1/1	0.0000
22	4	does	he	does	4/12	1/4	0.1666...
4	-1	does	his	best	4/12	2/4	0.1666...
31	-2	does	his	worst			
50	-3	does	*	*		1/4	0.1666...
19	4	he	does	he	4/12	4/4	0.0000
1	-1	he	does	he			
28	-2	he	does	his			
47	-3	he	does	*			
9	2	his	best	he	2/12	1/2	0.0833...
36	-1	his	worst	he		1/2	0.0833...
40	1	worst	he	does	1/12	1/1	0.0000
Sum of frequencies: 12						Second-order word entropy:	0.6666....

Concordances and text retrieval

Readers will have noticed that the tables produced were in fact concordances of the input sample text, the numbers of the leftmost column of which referenced the text by character. Further, the table used for the calculation of the second-order word entropy provides a handy structure for data retrieval. Say one wants to retrieve portions of text in which "he" and "best" occur in any order within n characters of each other. The table readily provides the information that "best" occurs only once, in position 13, and "he" four times in positions 19, 1, 28, and 47, from which the portions of text fulfilling the search conditions are easily identified. The retrieval process is considerably speeded up if identical word sequences have been further sorted by first character position, thus:

	Frequency	Pointing to:	
Pos.	or offset	Context	Word
13	1	best	he
4	4	does	he
22	-1	does	his
31	-2	does	his
50	-3	does	*
1	4	he	does
19	-1	he	does
28	-2	he	does
47	-3	he	does
9	2	his	best
36	-1	his	worst
40	1	worst	he

Table for calculation of second-order word entropy and text retrieval

Word-retrieval when sorting into alphabetical order and later searching and counting is far from straightforward (uppercase letters translated to lowercase, punctuation removed), and would be rather expensive computationally, particularly when sorting. Further, when retrieving text, one may be more interested in locating portions of text where certain words occur within a given number of words (or next to each other) rather than within a number of characters.

Optimizing text retrieval for speed

Once again, the modifications to the data structure are fundamentally simple.

A) Build a list of the different words in the input text (having converted them first to upper (or lower) case), along with a list of all the different sequences of punctuation marks.

B) Sort those two lists into alphabetical order, giving two separate dictionaries.

C) Encode the input text, replacing each word and punctuation sequence by its position in the dictionary. To be able to reconstruct the original text precisely, add to each word-and-punctuation pair information about whether and how the word was capitalized in the original, viz.:

1) all lowercase letters
2) all uppercase
3) uppercase initial, the rest lowercase.[4]

The encoded corpus now consists of fixed-length records, each representing a word, the punctuation which follows, and whether and how the word is capitalized.

D) Produce a list of all n-word sequences precisely in the same manner as that described earlier on, but this time using pointers to the initial 3-tuple of each sequence in the coded corpus, instead of pointers to the initial characters in the original text.

E) Sort that list into alphabetical order. The sorting process is considerably simpler and faster since words and punctuation are now represented in the source text by their alphabetical-order ranking, and word capitalization is now dissociated from the word itself.

For instance taking the sample text used in our earlier example:

"He does his best, he does, he does his worst, he does. * *"

[4] A word such as "MacMillan" is stored as two words ("Mac" and "Millan") separated by a null punctuation string.

A) Build two lists, one of its words and one of its punctuation strings:

Words:	he	does	his	best	worst	<null>
Punctuation:	<space>	<comma space>	<full stop>			

B) Sort them into alphabetical order[5]

Rank:	0	1	2	3	4	5
Words:	<null>	best	does	he	his	worst
Punctuation:	<none>	<space>	<comma space>	<full stop>		

C) Encode the input text as 3-tuples x:y:z in which x is the alphabetical rank of the word, y of the punctuation, and z the word's capitalization (say for instance 0 for all lowercase, 1 for uppercase initial, 2 for all uppercase):

Position:	1	2	3	4	5	6	7
Original:	He	does	his	best, he	does, he		
Translation:	3:1:1	2:1:0	4:1:0	1:2:0	3:1:0	2:2:0	3:1:0
Position:	8	9	10	11	12	13	14
Original:	does	his	worst,	he	does	<null>	<null>
Translation:	2:1:0	4:1:0	5:2:0	3:1:0	2:3:0	0:0:0	0:0:0

[5] Words and punctuation strings are ranked here in increments of 1 for simplicity's sake. But readers who have used or seen the MONKEY program demonstrated will have noticed that, in corcordance/text retrieval mode, it is capable of finding the closest approximation to a query containing a word that does not occur in the corpus. This is done very simply by ranking words in increments of 2. The words contained in the text retrieval query are first located in the dictionary using a binary search. When a word is not found it is encoded by the rank in-between. Imagine for instance that "her best" is to be retrieved from the text used in the example. "Best" is found in the dictionary, and encoded by its rank: 2. But the binary search fails for "her", returning the last two positions searched: "his" and "he", of rank 8 and 6. "Her" is therefore encoded by 7, and the corpus searched for a sequence 7,2.

D) List all sequences of n 3-tuples (here n=3):

Pos.	Pointing to			(Plain text:)
1	3:1:1	2:1:0	4:1:0	He does his
2	2:1:0	4:1:0	1:2:0	does his best,
3	4:1:0	1:2:0	3:1:0	his best, he
4	1:2:0	3:1:0	2:2:0	best, he does,
5	3:1:0	2:2:0	3:1:0	he does, he
6	2:2:0	3:1:0	2:2:0	does, he does
7	3:1:0	2:2:0	4:1:0	he does his
8	2:2:0	4:1:0	5:2:0	does his worst,
9	4:1:0	5:2:0	3:1:0	his worst, he
10	5:2:0	3:1:0	2:3:0	worst, he does.
11	3:1:0	2:3:0	0:0:0	he does.<null>
12	2:3:0	0:0:0	0:0:0	does.< null><null>

E) Sort into alphabetical order and compute frequencies or offsets as explained earlier:

Pos.	Frequency or offset	Context	(Plain text:)	Word
4	1	best	he	does
6	4	does	he	does
2	-1	does	his	best
8	-2	does	his	worst
12	-3	does	*	*
5	4	he	does	he
1	-1	he	does	his
7	-2	he	does	his
11	-2	he	does	*
3	2	his	best	he
9	-1	his	worst	he
10	1	worst	he	does

Given the above data structure, text generation, the computation of the word entropy to any order up to the length to which word sequences have been sorted, the production of concordances, and text retrieval are carried out essentially in the manner already described.

Applications

Almost 30 years ago Sukhotin proposed a number of algorithms for the automatic analysis of written texts. He once suggested that the "infinite-order entropy" of texts might provide a best objective function for such algorithms in general (Sukhotin 1973)[6]. The present algorithm brings the solution to the computation of the word entropy of texts to any arbitrary order, however high. As such it provides the objective function sought by Sukhotin.

The fact that the data structure at the basis of this algorithm is at the same time a concordance and a fast text-retrieval index, and that concordances are invaluable tools for semantic, syntactic and grammatical analysis, suggests to me that this very data structure will prove useful in yet to be discovered algorithms for minimizing or maximizing objective functions of the type discussed by Sukhotin.

Tretiakoff (1974) has produced evidence that, given two conflicting phonological interpretations of a text, the correct one showed a greater difference between its first and second-order character entropy. By analogy, we may surmise that, given two conflicting morphological interpretations, the correct one would show a greater difference between first and second-order word (or, rather here, morph) entropy.

A simple mental experiment lends credence to this conjecture.

Consider a text randomly segmented into pseudo-words.[7] The dictionary of pseudo-words will be very much larger than that of the words of the same text, correctly segmented. This means:

6 "Infinite-order" must be a *lapsus calami* or a translator's mistake since the nth-order entropy and beyond of a text 2n tokens long is necessarily zero.

7 The random segmentation should be such that the number of pseudo-words is the same (or at least roughly the same) as in the correctly segmented text.

1. The zeroeth-order entropy of an incorrect segmentation is higher than that of a correct segmentation.

2. Pseudo-words of an incorrect segmentation are predictable from their context to a far greater extent than is the case for a correct segmentation. Hence the nth-order entropy of an incorrectly segmented text will be far less than that of a correctly segmented one.

Therefore, the quantity H(n)-H(n-1) (in which H(n) denotes the nth order entropy) ought to be minimum in incorrect segmentations and maximum in correct segmentations.

That is precisely the phenomenon observed by Tretiakoff at the character level for n=2 when comparing conflicting phonological interpretations of texts where, for instance, the sequence "ng" was either left as such (to represent two phonemes) or replaced by a single character (one phoneme).

Apart from providing the objective function sought by Sukhotin in his attempts at developing what he termed "decipherment algorithms", the computation of word entropy of texts to any order ought to be useful in stylistic analysis, since while the zeroeth-order entropy is a measure of the richness of authors' vocabularies, higher-order values of the entropy measure their variety of expression as mirrored in their combining of words[8].

Readers will also have noticed that the encoding of the input text is likely to result in some degree of text compression. In fact, it once occurred to me that a solution to the segmentation of continuous text into its constituent morphs (a problem tackled by Sukhotin) should provide high ratios of text compression. To verify this hypothesis I carried out the following experiment:

1. Take a sample text — in this case the first 1000 lines of Leviticus, King James's version, as widely availible on diskette in the public domain. Its size is 52,388 bytes.

2. Compress it using LHARC, one of the more efficient standard data compression algorithms. The compressed file is 16,388 bytes long.

8 Thus it is futile to speak of "the word entropy of English" per se (or of any other language, for that matter).

3. Compute the first order word entropy of the original 1000 lines of text using the MONKEY program with option Capitalize OFF and all printing characters being declared as active letters. The text, with words thus defined, contains 9657 words (1540 different words), and its first order word entropy is 7.93756.

Therefore, using a static Huffman code, that text, as segmented, ought to be compressible to 7.93756x9657 = 76,653 bits approximately, i.e. 9582 bytes[9]. Certainly, it can be argued that the list of the different words encountered, that is, the dictionary of the text, would have to be transmitted first. Nevertheless:

1. The dictionary is sorted in alphabetical order, so that when two successive words share n initial characters, the first n characters of the second word need not be transmitted, only the value of n.

2. If a frequency count of the characters of the words in the dictionary is carried out first, it can be greatly compressed, again using static Huffman encoding.

3. The definition of a word used in this experiment (any continuous string of printable characters) is linguistically incorrect and quite sub-optimal. A better definition, or the output from an algorithm capable of accurately segmenting continuous text into its constituent morphs, ought to provide better compression ratios.

4. Finally, there is no reason why a protocol defining a set of default Huffman trees could not be developed. In which case it would suffice to transmit a few bytes specifying the default tree used.

[9] Approximately, since the lengths of the branches of a Huffman tree are necessarily integer values.

BIBLIOGRAPHY

Dewdney, Alexander K. (1989), "Computer Recreations." *Scientific American*, June issue.

Guy, Jacques B.M. (1990), "Fast high-order monkeys and a fast algorithm for calculating high-order character entropies". *Glottometrika* 12, 125-130. Bochum: Brockmeyer.

Shannon, C.E., Weaver, W. (1949), *The Mathematical Theory of Communication*. Urbana, Chicago, London: University of Illinois Press.

Sukhotin, B.V. (1973), "Méthode de déchiffrage, outil de recherche en linguistique". *T.A. Informations*, No.2.

Tretiakoff, A. (1974), "Identification de phonèmes dans la langue écrite par un critère d'information maximum". *Papers in Computational Linguistics* (Ferenc Papp & György Szépe, eds). Paris: Mouton, The Hague.

Rieger, Burghard (ed.)
Glottometrika 13
Bochum: Brockmeyer 1992, 186-195

Confidence limits for the entropies of phoneme frequencies

Peter Zörnig, Ursula Rothe

1. In Zörnig/Altmann (1984) a model has been presented to calculate the entropy of phoneme frequencies as a function of the number K of phonemes. The model was confirmed by computing the concerning sum of squared deviations.

Additional confirmations of the applicability of the model have been given in Rothe/Zörnig (1989) by computations of new language data from German and French. All examinations have shown that the prediction of the theoretical entropy by means of the phoneme number K of a language and the approximation formula of Zörnig/Altmann (1984) yielded good fittings.

The aim of the present contribution is to derive the theoretical variance and to prove that the observed entropy-values for 71 languages (cf. Zörnig/Altmann 1983 and 1984, Hug 1979) lie within the confidence limits of the theoretical entropy curve.

2. On the basis of the assumption that the rank-frequency distribution of phonemes obeys the Zipf-Mandelbrot law and using approximations for finite sums, Zörnig/Altmann (1984, p. 42, equation (7)) derived the formula

$$H_K = \mathrm{ld}\, e \ \ln\left(\sqrt{B+K)\ (B+1)}\ \ln\frac{B+K}{B+1}\right) \tag{1}$$

for the theoretical entropy H_K.

Here K denotes the number of phonemes in a language and B is a parameter. It was found that a fit using $B = 0{,}27$ or $B = 0{,}61$ is very good.

We assume that H_K is normally distributed for each point K, i.e. $HK \sim N(H,\sigma_K)$, where H and σ_k are the mean and the standard deviation of H_K. In order to derive the confidence limits for the observed entropies HK we consider the function

$$T = \hat{H}_K - J_K \tag{2}$$

Under the above assumption $T/\sigma(H_K - H_K)$ is normally distributed $(T/\sigma(\hat{H}_K\text{-}H_K) \sim N(0,1))$, i.e.

$$\frac{\hat{H}_K - H_K}{\sqrt{V(\hat{H}_K - H_K)}} = z \tag{3}$$

and the confidence limits are given by

$$|\hat{H}_K| \leq H_K \pm z\ \sqrt{V(\hat{H}_K - H_K)} \tag{4}$$

3. In order to compute the limits (4) we have to derive a formula for the variance

$$V(\hat{H}_K - H_K) = V(\hat{H}_K) + V(H_K) \tag{5}$$

a) The first term of the sum in (5) is given by

$$V(\hat{H}_K) = \frac{\sum_K \sum_{i=1}^{n_K} (\hat{H}_{K,i} - \overline{H}_K)^2}{\sum_K (n_K - 1)} \tag{6}$$

where $\hat{HK}_{,i}$ denotes the observed entropy of the i-th language with K phonemes, n_K is the number of languages with K phonemes and HK is given by

$$\overline{H}_K = \frac{1}{n_K} \sum_{i=1}^{n_K} \hat{H}_{K,i}\,. \tag{7}$$

b) In order to compute $V(H_K)$ in (5) we consider the Taylor-series in the environment of $S = E(K)$:

$$H_K = H_S + H'_S\,(K-S) \tag{8}$$

From $H_S = H_{E(K)} = E(H_K)$ and (8) follows

$$H_K - E(H_K) = H'_S\,(K-S) \quad \Rightarrow$$

$$V(H_K) = E(H_K - E(H_K)^2 = H'^2_S\,E(K-S)^2 \quad \Rightarrow$$

$$V(H_K) = H'^2_S\,V(K). \tag{9}$$

The first derivative of H_K is

$$H_K' = \left(\mathrm{ld}\,e \ln\left[\sqrt{(B+K)\,(B+1)}\,\ln\frac{B+K}{B+1}\right]\right)' \quad \Rightarrow$$

$$H_K' = \frac{\mathrm{ld}\,e}{\sqrt{(B+K)\,(B+1)}\,\ln\frac{B+K}{B+1}}\left(\sqrt{(B+K)\,(B+1)}\,\ln\frac{B+K}{B+1}\right)' \quad \Rightarrow$$

$$H_K' = \frac{\mathrm{ld}\,e}{\sqrt{(B+K)\,(B+1)}\,\ln\frac{B+K}{B+1}}\left(\frac{B+1}{2\sqrt{(B+K)\,(B+1)}}\ln\frac{B+K}{B+1} + \right.$$

$$\left. + \left(\sqrt{(B+K)\,(B+1)}\,\frac{(B+1)}{(B+K)\,(B+1)}\right)\right) \quad \Rightarrow$$

$$H_K' = \frac{\mathrm{ld}\,e}{(B+K)\ln\frac{B+K}{B+1}}\,(1 + \tfrac{1}{2}\ln\frac{B+K}{B+1}). \tag{10}$$

Inserting (10) in (9) yields

$$V(H_K) = V(K)\left[\frac{\mathrm{ld}\,e}{(B+S)\ln\frac{B+S}{B+1}}\,(1 + \tfrac{1}{2}\ln\frac{B+S}{B+1})\right]^2, \tag{11}$$

where the estimations

$$S = E(K) = 30.32\,,\; V(K) = 89.3 \tag{12}$$

(cf. Lehfeldt (1975)) can be used.

4. Inserting (5) in (4) yields

$$|\hat{H}_K| \le H_K \pm z\sqrt{V(\hat{H}_K) + V(H_K)} \tag{13}$$

The variances in (13) can be computed by (6) and (11) (cf. (12)).

The entropies $\hat{H}_K$, H_K, $\overline{H}_K$ and the confidence limits (cf. (13)) are presented in Table 1 for $B = 0.27$ and in Table 2 for $B = 0.61$ ($z = 1.96$).

The confidence band is shown in Figure 1 ($B = 0.27$) and Figure 2 ($B = 0.61$) respectively.

Tab.1. Entropies and confidence limits of 71 language samples ($B = 0.27$)

No.	Language	K	$\hat{H}_K$	H_K	$\overline{H}_K$	upper limit	lower limit
1	Hawaiian	13	3.370800	3.267970	3.304655	4.031234	2.504706
2	Hawaiian L	13	3.238510	3.267970	3.304655	4.031234	2.504706
3	Samoan	15	3.475850	3.453060	3.475850	4.216324	2.689796
4	Hawaiian	18	3.567110	3.682920	3.567110	4.446184	2.919656
5	Philipino	21	3.820320	3.872600	3.842963	4.635864	3.109336
6	Philipino	21	3.725080	3.872600	3.842963	4.635864	3.109336
7	Kaiwa	21	3.947760	3.872600	3.842963	4.635864	3.109336
8	Sea-Dayak L	21	3.878690	3.872600	3.842963	4.635864	3.109336
9	Estonian L	23	4.086150	3.982710	4.086150	4.745974	3.219446
10	Swahili L	24	4.023960	4.033790	3.993680	4.797054	3.270526
11	French L	24	3.963400	4.033790	3.993680	4.797054	3.270526
12	Albanian L	25	4.217630	4.082530	4.120430	4.845794	3.319266
13	Indonesian L	25	3.928940	4.082530	4.120430	4.845794	3.319266
14	Chamorro	25	4.214720	4.082530	4.120430	4.845794	3.319266
15	Dutch L	26	4.085840	4.129120	4.150465	4.892384	3.365856
16	English L	26	4.215090	4.129120	4.150465	4.892384	3.365856

No.	Language	K	$\hat{H}_K$	H_K	$\bar{H}_K$	upper limit	lower limit
17	Rumanian	27	4.253960	4.173760	4.066850	4.937024	3.410496
18	Spanish L	27	4.023920	4.173760	4.066850	4.937024	3.410496
19	Haussa L	27	3.922670	4.173760	4.066850	4.937024	3.410496
20	Dutch L	28	4.048850	4.216580	4.048850	4.979844	3.453316
21	Serbocroatian L	29	4.287100	4.257730	4.224410	5.020994	3.494466
22	Bulgarian L	29	4.219320	4.257730	4.224410	5.020994	3.494466
23	German L	29	4.172740	4.257730	4.224410	5.020994	3.494466
24	Indonesian	29	4.094240	4.257730	4.224410	5.020994	3.494466
25	Indonesian L	29	4.348650	4.257730	4.224410	5.020994	3.494466
26	German L	30	4.147520	4.297320	4.217357	5.060584	3.534056
27	Gujarati	30	4.497800	4.297320	4.217357	5.060584	3.534056
28	Italian L	30	4.006750	4.297320	4.217357	5.060584	3.534056
29	Italian	31	4.251220	4.335480	4.416613	5.098744	3.572216
30	Ukrainian L	31	4.565020	4.335480	4.416613	5.098744	3.572216
31	Russian L	31	4.433600	4.335480	4.416613	5.098744	3.572216
32	American English	32	4.487370	4.372290	4.481852	5.135554	3.609026
33	Hungarian	32	4.508420	4.372290	4.481852	5.135554	3.609026
34	Hungarian L	32	4.544990	4.372290	4.481852	5.135554	3.609026
35	Khasi	32	4.468860	4.372290	4.481852	5.135554	3.609026
36	Latvian L	32	4.428230	4.372290	4.481852	5.135554	3.609026
37	Russian L	32	4.453240	4.372290	4.481852	5.135554	3.609026
38	German	33	4.443530	4.407860	4.363648	5.171124	3.644596
39	Georgian	33	4.293130	4.407860	4.363648	5.171124	3.644596
40	Georgian	33	4.310650	4.407860	4.363648	5.171124	3.644596
41	Ostyak	33	4.375790	4.407860	4.363648	5.171124	3.644596
42	Ostyak	33	4.395140	4.407860	4.363648	5.171124	3.644596
43	Ostyak	34	4.406490	4.442240	4.398715	5.205504	3.678976
44	Ostyak	34	4.390940	4.442240	4.398715	5.205504	3.678976
45	French (HUG1)	35	4.647750	4.475530	4.614872	5.238794	3.712266
46	French (HUG2)	35	4.648450	4.475530	4.614872	5.238794	3.712266
47	French (HUG3)	35	4.679310	4.475530	4.614872	5.238794	3.712266
48	French (HUG4)	35	4.708380	4.475530	4.614872	5.238794	3.712266
49	French (HUG5)	35	4.709940	4.475530	4.614872	5.238794	3.712266
50	Czech	35	4.700600	4.475530	4.614872	5.238794	3.712266
51	Czech L	35	4.722750	4.475530	4.614872	5.238794	3.712266

No.	Language	K	$\hat{H}_K$	H_K	$\bar{H}_K$	upper limit	lower limit
52	French	35	4.101790	4.475530	4.614872	5.238794	3.712266
53	Marathi	38	4.514400	4.569420	4.688830	5.332684	3.806156
54	Bengali	38	4.863260	4.569420	4.688830	5.332684	3.806156
55	Hungarian	39	4.602810	4.598910	4.582193	5.362174	3.835646
56	English	39	4.709800	4.598910	4.582193	5.362174	3.835646
57	Armenian L	39	4.433970	4.598910	4.582193	5.362174	3.835646
58	German (HUG6)	40	4.755840	4.627570	4.759984	5.390834	3.864306
59	German (HUG7)	40	4.756760	4.627570	4.759984	5.390834	3.864306
60	German (HUG8)	40	4.767350	4.627570	4.759984	5.390834	3.864306
61	Russian	41	4.825690	4.655480	4.825690	5.418744	3.892216
62	Polish	42	4.725240	4.682630	4.821635	5.445894	3.919366
63	English	42	4.918030	4.682630	4.821635	5.445894	3.919366
64	Gujarati L	43	4.664620	4.709090	4.664620	5.472354	3.945826
65	English	44	4.906420	4.734890	4.819125	5.498154	3.971626
66	Slovak	44	4.731830	4.734890	4.819125	5.498154	3.971626
67	Swedish	45	4.840580	4.760050	4.840580	5.523314	3.996786
68	Ukrainian	46	4.627950	4.784600	4.627950	5.547863	4.021336
69	Hindi	52	4.584600	4.920670	4.584600	5.683934	4.157406
70	Burmese L	68	5.230640	5.213390	5.230640	5.976654	4.450126
71	Vietnamese	74	5.160550	5.304330	5.160550	6.067594	4.541066
		obs. var.: 0.019967 theor. var.: 0.131681					
		variance in (5): 0.151648					

Tab.2. Entropies and confidence limits of 71 language samples ($B = 0.61$)

No.	Language	K	$\hat{H}_K$	H_K	$\bar{H}_K$	upper limit	lower limit
1	Hawaiian	13	3.370800	3.320770	3.304655	2.545417	4.096123
2	Hawaiian L	13	3.238510	3.320770	3.304655	2.545417	4.096123
3	Samoan	15	3.475850	3.509490	3.475850	2.734137	4.284843
4	Hawaiian	18	3.567110	3.743820	3.567110	2.968467	4.519173
5	Philipino	21	3.820320	3.937150	3.842963	3.161797	4.712503
6	Philipino	21	3.725080	3.937150	3.842963	3.161797	4.712503
7	Kaiwa	21	3.947760	3.937150	3.842963	4.712503	3.161797
8	Sea-Dayak L	21	3.878690	3.937150	3.842963	4.712503	3.161797

No.	Language	K	$\hat{H}_K$	H_K	$\bar{H}_K$	upper limit	lower limit
9	Estonian L	23	4.086150	4.049350	4.086150	4.824703	3.273997
10	Swahili L	24	4.023960	4.101390	3.993680	4.876743	3.326037
11	French L	24	3.963400	4.101390	3.993680	4.876743	3.326037,
12	Albanian L	25	4.217630	4.151040	4.120430	4.926393	3.375687
13	Indonesian L	25	3.928940	4.151040	4.120430	4.926393	3.375687
14	Chamorro	25	4.214720	4.151040	4.120430	4.926393	3.375687
15	Dutch L	26	4.085840	4.198510	4.150465	4.973863	3.423157
16	English L	26	4.215090	4.198510	4.150465	4.973863	3.423157
17	Rumanian	27	4.253960	4.243970	4.066850	5.019323	3.468617
18	Spanish L	27	4.023920	4.243970	4.066850	5.019323	3.468617
19	Haussa L	27	3.922670	4.243970	4.066850	5.019323	3.468617
20	Dutch L	28	4.048850	4.287580	4.048850	5.062933	3.512227
21	Serbocroatian L	29	4.287100	4.329490	4.224410	5.104843	3.554137
22	Bulgarian L	29	4.219320	4.329490	4.224410	5.104843	3.554137
23	German L	29	4.172740	4.329490	4.224410	5.104843	3.554137
24	Indonesian	29	4.094240	4.329490	4.224410	5.104843	3.554137,
25	Indonesian L	29	4.348650	4.329490	4.224410	5.104843	3.554137
26	German L	30	4.147520	4.369810	4.217357	5.145163	3.594457
27	Gujarati	30	4.497800	4.369810	4.217357	5.145163	3.594457
28	Italian L	30	4.006750	4.369810	4.217357	5.145163	3.594457
29	Italian	31	4.251220	4.408660	4.416613	5.184013	3.633307,
30	Ukrainian L, 31	4.5	4.565020	4.408660	4.416613	5.184013	3.633307
31	Russian L	31	4.433600	4.408660	4.416613	5.184013	3.633307
32	American English	32	4.487370	4.446140	4.481852	5.221493	3.670787
33	Hungarian	32	4.508420	4.446140	4.481852	5.221493	3.670787
34	Hungarian L	32	4.544990	4.446140	4.481852	5.221493	3.670787
35	Khasi	32	4.468860	4.446140	4.481852	5.221493	3.670787
36	Latvian L	32	4.428230	4.446140	4.481852	5.221493	3.670787
37	Russian L	32	4.453240	4.446140	4.481852	5.221493	3.670787
38	German	33	4.443530	4.482340	4.363648	5.257693	3.706987
39	Georgian	33	4.293130	4.482340	4.363648	5.257693	3.706987
40	Georgian	33	4.310650	4.482340	4.363648	5.257693	3.706987
41	Ostyak	33	4.375790	4.482340	4.363648	5.257693	3.706987
42	Ostyak	33	4.395140	4.482340	4.363648	5.257693	3.706987
43	Ostyak	34	4.406490	4.517340	4.398715	5.292693	3.741987

No.	Language	K	$\hat{H}_K$	H_K	$\bar{H}_K$	upper limit	lower limit
44	Ostyak	34	4.390940	4.517340	4.398715	5.292693	3.741987
45	French (HUG1)	35	4.647750	4.551220	4.614872	5.326573	3.775867
46	French (HUG2)	35	4.648450	4.551220	4.614872	5.326573	3.775867
47	French (HUG3)	35	4.679310	4.551220	4.614872	5.326573	3.775867
48	French (HUG4)	35	4.708380	4.551220	4.614872	5.326573	3.775867
49	French (HUG5)	35	4.709940	4.551220	4.614872	5.326573	3.775867,
50	Czech	35	4.700600	4.551220	4.614872	5.326573	3.775867
51	Czech L	35	4.722750	4.551220	4.614872	5.326573	3.775867
52	French	35	4.101790	4.551220	4.614872	5.326573	3.775867
53	Marathi	38	4.514400	4.646770	4.688830	5.422123	3.871417
54	Bengali	38	4.863260	4.646770	4.688830	5.422123	3.871417
55	Hungarian	39	4.602810	4.676780	4.582193	5.452133	3.901427
56	English	39	4.709800	4.676780	4.582193	5.452133	3.901427
57	Armenian L	39	4.433970	4.676780	4.582193	5.452133	3.901427
58	German (HUG6)	40	4.755840	4.705650	4.759984	5.481003	3.930297
59	German (HUG7)	40	4.756760	4.705650	4.759984	5.481003	3.930297
60	German (HUG8)	40	4.767350	4.7056650	4.759984	5.481003	3.930297
61	Russian	41	4.825690	4.734330	4.825690	5.509683	3.958977
62	Polish	42	4.725240	4.761960	4.821635	5.537313	3.986607
63	English	42	4.918030	4.761960	4.821635	5.537313	3.986607
64	Gujarati L	43	4.664620	4.788880	4.664620	5.564233	4.013527
65	English	44	4.906420	4.815110	4.819125	5.590463	4.039757
66	Slovak	44	4.731830	4.815110	4.819125	5.590463	4.039757
67	Swedish	45	4.840580	4.840700	4.840580	5.616053	4.065347
68	Ukrainian	46	4.627950	4.865670	4.627950	5.641023	4.090317
69	Hindi	52	4.584600	5.004010	4.584600	5.779363	4.228657
70	Burmese L	68	5.230640	5.301440	5.230640	6.076793	4.526087
71	Vietnamese	74	5.160550	5.393800	5.160550	6.169153	4.618447
		obs. var.: 0.019967 theor. var.: 0.136523 variance in (5): 0.156190					

Fig. 1. Confidence band of the entropies of the phoneme frequencies with 71 language samples ($B = 0.27$)

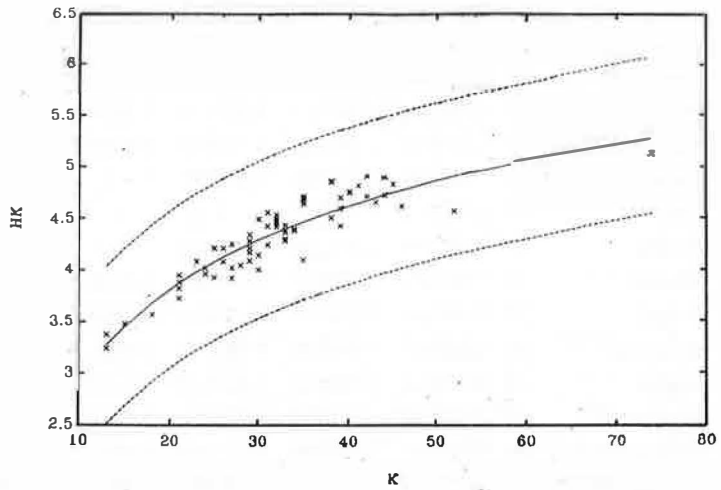
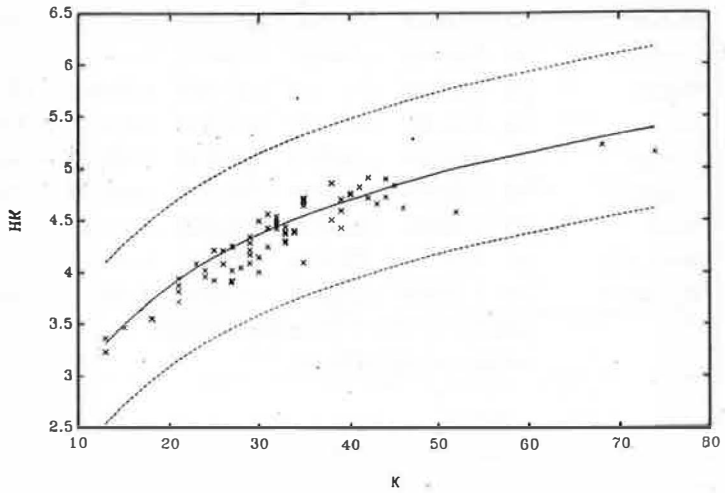


Fig. 2. Confidence band of the entropies of the phoneme frequencies with 71 language samples ($B = 0.61$)



Bibliography

Hug, Marc (1979), *La distribution des phonèmes en français. Die Phonemverteilung im Deutschen.* Genève: Slatkine.

Lehfeldt, Werner (1975), "Die Verteilung der Phonemzahl in den natürlichen Sprachen". *Phonetica 31*, 274-287.

Rothe, Ursula, Zörnig, Peter (1990), "The entropy of phoneme frequencies. German and French". *Glottometrika 11*.

Zörnig, Peter, Altmann, Gabriel (1983), "The Repeat Rate of Phoneme Frequencies and the Zipf-Mandelbrot Law". *Glottometrika 5*, 205-211. Bochum: Brockmeyer.

Zörnig, Peter, Altmann, Gabriel (1984), "The entropy of phoneme frequencies and the Zipf-Mandelbrot law". *Glottometrika 6*, 41-47. Bochum: Brockmeyer.

Rieger, Burghard (ed.):
Glottometrika 13
Bochum: Brockmeyer 1992, 196-204

Towards an objective function for the automatic phonemicization of texts transcribed phonetically?

Jacques B.M. Guy

Overview

Sukhotin (1962) has proposed an algorithm that, given a text in phonemic or near-phonemic transcription, very accurately classifies its symbols into vow-els and consonants (for later research on this subject see Boy 1977). At about the same time Sukhotin had proposed a number of algorithms for the automatic analysis of natural-language texts, amongst which are of particular interest here:

1. An algorithm for the segmentation of continuous text into its constituent morphemes.
2. An algorithm for discovering the pronunciation of letters, given texts transcribed in phonemic or near-phonemic writing systems.

Sukhotin candidly remarked that in some cases the results obtained were at first "incredibly bad", but that experimenting with diverse objective functions brought some improvements. Sukhotin's work, alas, was completely hindered by the lack of computing and programming facilities at the time. Only his simplest algorithms were ever run on computers. The others, too complex to be programmed then, had to be tested by hand, an exceedingly time-consuming task that prevented proper experimenting and development. Apart from the topic of Boy's 1977 Ph.D. thesis, Sukhotin's works seem to have fallen into oblivion.

Sukhotin's 1962 original vowel-recognition algorithm is one of the few for which he does not define an objective function.

Many years later Tretiakoff (1976) produced evidence that the difference between first and second-order character entropy is maximized in texts where each letter has been correctly replaced by "C" (consonant) or "V" (vowel), thus unwittingly providing Sukhotin's missing objective function (Tretiakoff was not aware of Sukhotin's works).

In the following article theoretical and experimental evidence is brought that the difference between first and second-order character entropy of texts transcribed phonemically provides a measure of the correctness of the phonemic analysis at the basis of the transcription system.

Hypothesis

Since

1. a phonemic transcription of a text is essentially a phonetic rendering from which the redundant phonological features of the language have been abstracted;

2. phonemes are identified by examining the distribution and co-occurrence restrictions of their putative allophones, that is, by examining their phonetic context;

it follows that the difference between first and nth-order character entropy is a measure of the amount of redundancy removed when contexts of n-1 phonetic symbols are taken into account in the phonemic analysis of a sample text.[1]

Therefore, given several conflicting phonemic analyses of a language and a sample text in that language transcribed according to those conflicting analyses, we should expect the transcription with the greatest difference between first and

[1] The value chosen for n, however, should be kept quite low: first, allophonic variations are seldom conditioned by anything but the immediate phonetic environment; second, the frequency distribution of strings of 4 characters and longer describes, at least for most languages, lexical rather than phonological properties.

second-order (and perhaps also between first and third-order) character entropy to correspond to the most correct phonemic interpretation.

Experiment

The phonetic transcriptions of the first three fables ("P'ostrye ovtsy", "Or'ol i kury", and "Mor'ak") in Boyanus 1967 were input into a computer file.[2] Six phonemic and mock-phonemic versions of the original file ("RUS.0") were then produced, replacing every occurrence of selected characters by another.

File "RUS.1": The generally accepted interpretation of the vowel system of Russian:

@ ([æ]) → a

9 ([ə]) → a

1 ([ɪ]) → i

w ([i̯]) → i

& ([ɛ]) → e

y ([ü]) → u

---

2 Russian was chosen over English because:
1. there is general agreement about its phonology, which there is not for English,
2. the interpretation of palatalization as either a vocalic or consonantal feature made for an interesting experiment.

Phonetic symbols not available on the standard computer keyboard were represented as follows:
[æ]= @ [ɛ]= & [ɪ]= 1 [i̯]= w [ü]= y
[ʒ]= 3 [ʃ]= $ [ʃ ']= 4 [ʧ]= q [ts]= c [ə]= 9
The entropy was computed using the algorithm described in Glottometrika 12, with spaces disregarded (option "Spaces OFF"), lowercase letters left as such (option "Capitalize OFF"), and the alphabet of active letters declared as:
1 2 3 4 5 6 7 8 9 0
@ $ & '
a b c d e f g h i j k l m n o p q r s t u v w x y z
ABCDEFGHIJKLMNOPQRSTUVWXYZ

File "RUS.2": An incorrect interpretation:

@ ([æ]) → i

9 ([ə]) → e

1 ([ɪ]) → a

w ([i̯]) → e

& ([ɛ]) → u

y ([ü]) → a

File "RUS.3": A grossly incorrect interpretation:

@ ([æ]) → t

9 ([ə]) → m

1 ([ɪ]) → p

w ([i̯]) → k

& ([ɛ]) → k

y ([ü]) → p

File "RUS.4": A slightly incorrect interpretation:

a ([a]) → o

9 ([ə]) → o

1 ([ɪ]) → i

w ([i̯]) → i

& ([ɛ]) → e

y ([ü]) → u

File "RUS.5": Same as file "RUS.1", but with palatalization merged with vowels, e.g.: 'a → A, 'e → E, etc.

File "RUS.6": Same as file "RUS.1", but with palatalization merged with consonants, e.g.: b' → B, d' → D, etc.

The results obtained are summarized in the table hereunder (higher-order val-ues of the entropy are given only for curiosity's sake):

	RUS.0	RUS.1	RUS.2	RUS.3	RUS.4	RUS.5	RUS.6
Different characters	32	26	26	26	26	31	41
Total characters	3828	3828	3828	3828	3828	3436	3333
h0	5.00000	4.70044	4.70044	4.70044	4.70044	4.95420	5.35755
h1	4.56352	4.22935	4.21989	4.28692	4.10642	4.51855	4.63200
h2	3.47861	3.30374	3.37287	3.49004	3.20176	3.60223	3.60367
h3	2.32221	2.39795	2.45459	2.45253	2.42606	2.34389	2.27757
h4	1.00327	1.24882	1.19829	1.08839	1.35211	0.92995	0.86728
h5	0.34090	0.46598	0.42855	0.37950	0.52864	0.23627	0.22178
h6	0.11126	0.15752	0.13328	0.11891	0.18116	0.07273	0.06345
h1-h2	1.08491	0.92561	0.84702	0.79688	0.90466	0.91632	1.02833
h1-h3	2.24131	1.83140	1.76530	1.83439	1.68036	2.17366	2.35443

Only the figures in columns 2 to 5 (files "RUS.1" to "RUS.4"), obtained on texts with the same number of characters in their alphabets, and of the same length, are strictly comparable with one another. The difference between first and second-order character entropy is greatest (0.925613) for the correct transcription (file "RUS.1"), and least (0.79688) for the most aberrant transcription in which some vowel symbols were replaced by consonants (File "RUS.3"). The slightly incorrect transcription in file "RUS.4" scores the second highest difference (0.90466). There appears to be no known statistical tests of significance of values obtained for the character or word entropy of texts (or, at least, I have been unable to find any in the literature), so that it is only likely, but not absolutely certain, that the differences observed in this experiment are significant.

The figures for files "RUS.5" (palatalization as a vocalic feature) and "RUS.6" (palatalization as a consonantal feature) are interesting: the difference between first and second-order character entropy happens to be greater for the text in which palatalization was transcribed as a consonantal feature (the generally accepted modern interpretation). Those two texts, however, have different lengths (3436 vs 3333 characters) and alphabets (31 vs 41 characters), so that, again in the absence of an adequate statistical test, no certain conclusion can be drawn in this case.

Further Questions

Although the absence of statistical tests of significance for values of the character entropy of texts prevents us from drawing firm conclusions, the results of the experiment indicate a strong possibility that the difference between first and second-order character entropy of texts transcribed phonemically provides a measure, i.e. an objective function, of the correctness of a phonemic analysis. If it is indeed so, then there must exist an algorithm capable of maximizing this objective function, and, therefore, of automatically phonemicizing texts transcribed phonetically.

Tretiakoff has produced evidence that the difference between first and second-order character entropy is also a measure of the correctness of the partition of an alphabet into vowels and consonants, and this may appear to prevent this measure from being also used as an estimator of phonemic correctness.

But there is no incompatibility there. A phoneme is ultimately a set the members of which are its allophones (each often associated with one or several conditioning environments). A vowel (or a consonant) can likewise be con-strued as a set. The notion of phoneme as set is applicable to archiphonemes: an archiphoneme is a set, the members of which are phonemes, each possibly associated with conditioning environments.

The one and only fundamental difference between the phonemicization process and Sukhotin's vowel recognition algorithm is that the latter forces the partition of the alphabet of a text into exactly two subsets; the phonemicization process,

on the other hand, does not specify the number of subsets into which the alphabet is to be partitioned.[3]

In this light, phonemic analysis becomes a clustering process that ends in a binary classification (vowel vs consonant) where several layers of what we might perhaps call archiphonemes can intervene between phonemes proper and the vowel/consonant dichotomy.

The fact that Sukhotin's vowel-recognition algorithm is simple and computationally very inexpensive gives hope that simple, efficient algorithms for the automatic phonemic analysis of texts may be discovered. Automatic phonemic analysis, however, will require not only the partition of the alphabet of a text (as in the translation of file "RUS.0" into files "RUS.1" to "RUS.4"), but also the segmentation of continuous texts into phonological units (as in the translation of file "RUS.1" into files "RUS.5" and "RUS.6", in which the palatalization symbol was merged with the preceding or following letter). The latter problem is identical to that of the segmentation of continuous text into morphemes, for which Sukhotin has also proposed an algorithm.[4]

---

3 The analysis of the vowel system of the Sakao language (spoken in Espiritu Santo, New Hebrides — now Vanuatu) would likely profit from such an approach, which allows more than one layer of archiphonemes. One is forced to resort to vowel archiphonemes to account for extensive neutralizations in the twelve-vowel system of Sakao; yet the resulting analysis remains awkward and feels fundamentally unsatisfactory: one particular vowel archiphoneme, for instance, is defined for height only, its environment deciding whether it is back, front, rounded, or unrounded; another is back, rounded, with its height entirely conditioned by its environment; yet another has vocalicity for its only distinguishing feature, every other feature being determined by its environment (Guy 1974). Similar problems affect, but to a lesser extent, the consonant system of some languages of Oba (New Hebrides).

4 I implemented Sukhotin's text segmentation algorithm in 1977 and experimented on samples of English and Asmat (a language of Papua-New Guinea). The results were poor, about 65% of strings given as morphemes being true morphs or morphemes for the English sample, and 60% for the Asmat sample. The algorithm itself appeared to be linguistically correct and sensible. Experimenting with alternative objective functions brought an improvement of some five percentage points. Those disappointing results seem due to inadequate objective functions. Sukhotin mentioned successive-order values of the character or word entropy as a more likely efficacious objective function for his decipherment algorithms in general, but character entropy (and a fortiori word entropy)

Finally, we must note that the identification of the most correct phonemic transcription by using the difference between first and second-order character entropy is carried out without reference to the phonetic values represented by the characters of the text. It seems extraordinary, even impossible, that a correct phonemic interpretation could be identified without knowing what the symbols of the text actually represent. And yet, Sukhotin's vowel-recognition algorithm also functions in the absence of any data about the phonetic values of the characters of the input text.[5] One rational explanation would be that, in most and perhaps all languages, the distributional properties of phonological units are so strongly determined by their acoustic and/or articulatory features that a phonologically meaningful classification (i.e. phonemic analysis) can be obtained on the basis of those distributional properties alone.

REFERENCES

Boy, J. (1977), *Dechiffrierungsalgorithmen zur phonetischen Identifikation von Buchstaben*. Bochum: Brockmeyer.

Boyanus, S.C. (1967), *Russian Pronunciation and Russian Phonetic Reader*. London: Percy Lund, Humphries & Co.

Cherry, C., M. Halle, and R.O. Jakobson (1953), "Towards the Logical Description of Languages in their Phonemic Aspect." *Language* 29, 34-46.

Guy, J.B.M. (1974), *A Grammar of the Northern Dialect of Sakao*. Canberra: Pacific Linguistics.

---

5 Strictly speaking, however, Sukhotin's algorithm does not identify vowels (or consonants) as such. It only classifies the characters of a text into two groups, one of which contains vowels and the other consonants, and there is strictly no way of knowing which contains which. However, it so happens that the first character to be assigned to a group is the most frequent one, and Sukhotin labels that group "vowels". But there is no compelling reason why the most frequent phoneme of a language should not be a consonant. Nevertheless, and although I have tried Sukhotin's algorithm on a vast number of languages, I have not yet encountered one with a consonant for its most frequent phoneme.

Paducheva, E.V. (1963), "Measurement of Information and the Study of Language". In: *Exact Methods in Linguistic Research* (translated by David Hays and Dolores mohr), 119-180. Berkeley/Los Angeles: University of California Press.

Shannon, E.C. (1948), "A mathematical theory of communication." *Bell Technical System Journal*, Vol. 27, No.4, 379-423.

Sukhotin, B.V. (1962), "Eksperimental'noe vydelenie klassov bukv s pomoshch'ju elektronnoj vychislitel'noj mashiny." *Problemy strukturnoj lingvistiki* 234, 198-206.

Sukhotin, B.V. (1966), "Issledovanie jazyka deshifrovochnymi metodami." *Russkii jazyk v shkole* 6, 14-22.

Sukhotin, B.V. (1973), "Méthode de déchiffrage, outil de recherce en linguistique." *T.A. Informations*, 2, 3-43.

Trétiakoff, A. (1976), "Identification des phonèmes dans la langue écrite par un critère d'information maximum." *Papers in Computational Linguistics* (Ferenc Papp & György Szépe eds). Paris: Mouton, 403-408.

Rieger, Burghard (ed.)
Glottometrika 13
Bochum: Brockmeyer 1992, 205-218

The Zipf-Mandelbrot Law and the Interdependencies Between Frequency Structure and Frequency Distribution in Coherent Texts

Peter Zörnig, Moisei Boroda

0. Introduction

The modelling of the repetition organization of a coherent text, i.e. the determination of the relations between parameters such as text length, inventory size and the frequency of occurrence of certain text units is very important in text analysis. Indeed the question whether there are general regularities in the repetition organization, to what extent they are relevant for various languages, by which mathematical models these relations could be described, etc. fall within the framework of discovering the principles governing text generation. Repetition plays a particularly important role in music and poetry. Investigations carried out in this direction have revealed a number of regularities (cf. Altmann 1988; Arapov 1988; Boroda 1982, 1990; Orlov 1982; Tuldava 1987 a.o.), the most general of which is the fact that in every coherent text only a few units occur frequently, whereas large groups of units occur only once, twice, etc. When the frequencies of all the different units of a text are arranged in a (not necessarily strict) decreasing order, i.e. if we assign rank 1 to the most frequent unit, rank 2 to the second-most frequent unit etc., the frequency organization can be represented by a stair-step curve (Fig. 1).

The above-mentioned fact can be observed in Fig. 1 as follows:
The left part of the diagram consists essentially of steps of length 1, while

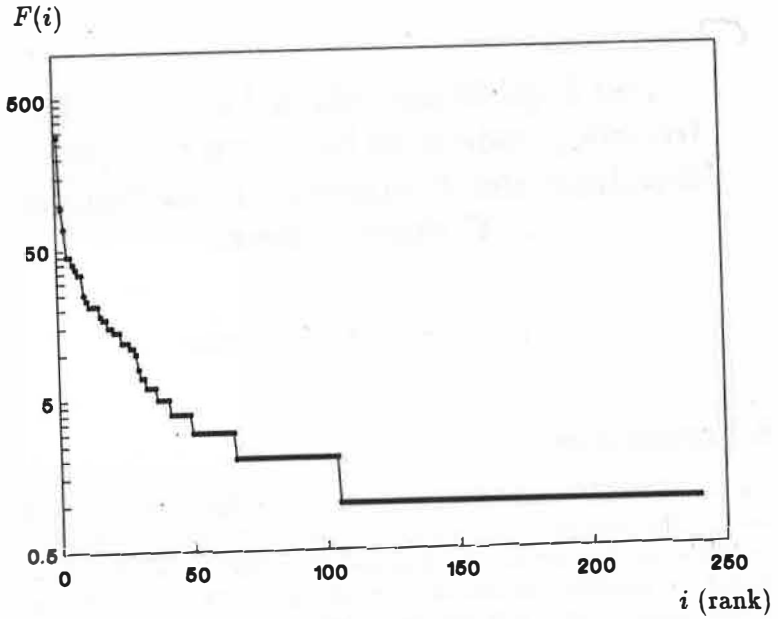


Fig. 1
Frequency structure of the melodic units F-motifs in the Prelude and Fugue WTC 1:20 of J.S. Bach.

in the right part there are mainly steps of length greater than one, whereas the length increases with increasing rank. Such a subdivision of the stairs into two parts has been possible for all (sufficiently large) texts investigated up to now. As a consequence, the relations between the frequencies of units were considered here from two points of view (cf. section 1 and Boroda/Zörnig 1990):

- Motivated by the *strictly* decreasing left part of the stairs, the rank i was assigned to the height of the steps, i.e. to the frequency F_i; we call this relation the frequency structure (FS).
- As is suggested by the right part of the diagram, to each frequency F the number V_F of different text units occurring F

Table 1
Models for the Frequency Distribution (FD)

Formulae are listed for the number V_F of different text units with frequency F (cf. (3)) (L := text length, V := inventory size)

a) Waring-Herdan Model (cf. Herdan 1966):

$$V_{F+1} = V_F \frac{a+F-1}{x+F},$$

where

$$a = \left(Q - \frac{V}{L} - 1\right)^{-1}$$

$$x = aQ$$

$$Q = \left(1 - \frac{V_1}{V}\right)^{-1}$$

b) the Krallmann Model (1966):

$$V_F = cF^{-k}e^{bF}$$

c) the Orlov Model (1982):

$$V_F = \frac{V}{F(F+1)}$$

d) the Hajtun Model (1983):

$$V_F = AF^{-(1+1/c)},$$

where c represents the parameter in the ZM-Law and $A \approx \frac{V}{c}$

e) the Krylov Model (1987) (cf. Krylov 1987, Tuldava 1987):

$$V_F = cF^{-2}e^{b/F}$$

where c and b are parameters dependent on the text length, the frequency of the most frequent unit and the vocabulary size.

times each, was assigned. The number V_F can be interpreted as length of a respective step. We will call this relation the frequency distribution (FD).

The "dual" representation of frequency relations led to a "dualism" in their description, too, i.e. as a rule the FS and the FD have been described by different models. While the former is generally described by the Zipf-Mandelbrot Law (ZM-Law, cf. section 1), there exist numerous other models for the FD (cf. Tab. 1).

In this article we prove that, assuming that FS follows the ZM-Law, FD also follows (at least approximately) the ZM-Law with other values of parameters; i.e. on the above assumption FS and FD can be described by *one and the same* model. We also show how the parameters of FD can be obtained by those of FS (section 2). In section 3 we illustrate the dependence between the parameter values using Goethe's "Erlkönig" as a text and discuss the adequacy of the model. Finally we investigate a slight modification of it (section 4).

1. Fundamentals

Since the expressions "frequency structure" (FS) and "frequency distribution" (FD) are often used with different meanings or in a "fuzzy" sense, we would like to develop an exact definition by means of a small hypothetical example. Consider the formal text

a b b c d e f d d e e f f
g g g h g h k h k k k k h, (1)

where each letter represents one of the investigated units of the text; (1) could represent for example a literary as well as a musical text. Now the FS of (1) is obtained if to each text unit i the frequency of occurrence F_i is assigned (ordered according to decreasing frequency):

text unit i	frequency F_i
k	5
g	4
h	4
d	3
e	3
f	3
b	2
a	1
c	1

(2)

From the FS (2) we obtain the FD as follows. To each frequency F we assign the number V_F of different text units with this frequency, i.e. the number of occurrences of this frequency:

frequency F	number of occurrences V_F
3	3
1	2
4	2
2	1
5	1

(3)

List (3) states that frequency 3 occurs 3 times, frequency 1 occurs 2 times and so on.

Obviously the following general relations hold for the numbers F_i and V_F:

$$L = \sum_{i=1}^{V} F_i,$$

$$V = \sum_{F=1}^{F_{max}} V_F,$$

where L, V, F_{max} denote the text length, the inventory size and the maximal frequency of text units, respectively. For the example above $L = 26$, $V = 9$, $F_{max} = 5$.

In the following we assume that the ZM-Law is valid for the FS of a given coherent text, i.e. that the i-th frequency F_i in (2) can be fitted by an expression of the form

$$F_i = \frac{Z}{(i+B)^C} \quad i = 1, \ldots, V, \tag{4}$$

where B, C are constants and Z is a standardizing constant, i.e.

$$Z = L\left(\sum \frac{1}{(i+B)^C}\right)^{-1}$$

2. The Interrelations Between FS and FD

We start our considerations with equation (4).

Since the frequencies in (4) must be integral, we can round the right side off to the nearest integer, i.e. we can modify (4) to

$$F_i = \left[\frac{Z}{(i+B)^C} + \frac{1}{2}\right], \tag{5}$$

where $[x]$ denotes the greatest integer $\leq x$. Now V_F in (3) is the cardinality of the set

$$\{i | F_i = F\},$$

i.e.

$$V_F = |\{i | F_i = F\}|.$$

This implies using (5)

$$\begin{aligned} V_F &= \left|\left\{i \left| \left[\frac{Z}{(i+B)^C} + \frac{1}{2}\right] = F\right.\right\}\right| \\ &= \left|\left\{i | F - \frac{1}{2} \leq \frac{Z}{(i+B)^C} < F + \frac{1}{2}\right\}\right| \\ &= \left|\left\{i \left| \left(\frac{Z}{F+\frac{1}{2}}\right)^{1/C} - B < i \leq \left(\frac{Z}{F-\frac{1}{2}}\right)^{1/C} - B\right.\right\}\right|, \end{aligned}$$

where the last equation is obtained by isolating i. From this follows

$$V_F = \left|\left\{i \left| \left[\left(\frac{Z}{F+\frac{1}{2}}\right)^{1/C} - B\right] < i \leq \left[\left(\frac{Z}{F-\frac{1}{2}}\right)^{1/C} - B\right]\right.\right\}\right|,$$

since i is an integer. The last expression yields

$$V_F = \left[\left(\frac{Z}{F-\frac{1}{2}}\right)^{1/C} - B\right] - \left[\left(\frac{Z}{F+\frac{1}{2}}\right)^{1/C} - B\right], \tag{6}$$

which can be used as an initial model for the FD.

Expression (6) represents a simple generalization of the numbers N_n in Haight (1969: 225, eq. (1)) which were used to derive the Zeta distribution and the harmonic distribution.

In the following we derive an approximation for the relatively complicated expression (6) which is also of the form (4) (cf. also (11)).

Omitting the "integral operators" [ ] in (6) yields the first approximation

$$V_F \approx \left(\frac{Z}{F-\frac{1}{2}}\right)^{1/C} - \left(\frac{Z}{F+\frac{1}{2}}\right)^{1/C}. \tag{7}$$

Using the definition

$$g(F) := \left(\frac{Z}{F}\right)^{1/C}, \tag{8}$$

(7) can be represented as

$$V_F \approx -\frac{g\left(F+\frac{1}{2}\right)-g\left(F-\frac{1}{2}\right)}{\left(F+\frac{1}{2}\right)-\left(F-\frac{1}{2}\right)}. \tag{9}$$

According to the mean value theorem there exists a $\hat{B}$, $|\hat{B}| < \frac{1}{2}$ such that

$$-\frac{\partial}{\partial F} g\left(F+\hat{B}\right)$$

equals the right side in (9)*.

This yields the relation

$$V_F \approx -\frac{\partial}{\partial F} g\left(F+\hat{B}\right). \tag{10}$$

Using (8) we obtain

$$V_F \approx \frac{Z^{1/C}}{C\left(F+\hat{B}\right)^{1+1/C}}. \tag{11}$$

With the notations

$$\left.\begin{aligned} \hat{Z} &:= \frac{1}{C} Z^{1/C}, \\ \hat{C} &:= 1+\frac{1}{C} \end{aligned}\right\} \tag{12}$$

we obtain

$$V_F \approx V_F^* := \frac{\hat{Z}}{\left(F+\hat{B}\right)^{\hat{C}}}. \tag{13}$$

* To be exact, the parameter $\hat{B}$ depends on F, but in the following approximation we assume that one and the same parameter $\hat{B}$ can be taken for all values of F.

The derivation above shows that the FD approximately obeys the ZM-Law with parameters $\hat{Z}$, $\hat{B}$, $\hat{C}$, when FS obeys the ZM-Law with parameters Z, B, C. The parameters $\hat{Z}$, $\hat{C}$ are given by (12), and $\hat{B}$ can be determined experimentally (cf. section 3).

In Boroda/Zörnig (1990) it is shown that model (13) can be well fitted to the empirical FD of certain musical texts.

3. Illustration and Fit to Empirical Data

In this section we give an example to illustrate the dependencies between FS and FD. We take Goethe's "Erkönig" as the text, where the word forms are considered as text units.

The FS is given in Tab. 2 (cf. Altmann 1988: 73, Tab. 2 13(a)).

Tab. 3 contains the corresponding FD (cf. (2), (3)).

The parameters $\hat{Z}$, $\hat{C}$ in Tab. 2 are computed by (12):

$$\hat{Z} = \frac{1}{C} Z^{1/C} = \frac{1}{0.709} 23.175^{(1/0.709)} = 118.7$$

$$\hat{C} = 1+\frac{1}{C} = 1+\frac{1}{0.709} = 2.410.$$

The parameter $\hat{B}$ was chosen such that V_1^* equals the corresponding empirical value, i.e. $\hat{B} = 0.15$.

Because the parameter $\hat{Z}$ in (13) is not a standardizing constant, the sums of empirical and theoretical values for the FD can be different. Though this difference is small in our example (cf. Tab. 3) we cannot apply the chi-square test in general to test the goodness of fit. Instead we compute the determination coefficient

$$\mathbf{Krit} = 1 - \frac{\sum_{F=1}^{11} (V_F - V_F^*)^2}{\sum_{F=1}^{11} \left(V_F - \hat{V}\right)^2}, \tag{14}$$

Table 2
FS of the word forms in Goethe's "Erlkönig"

Word forms	Frequencies of word forms	
	empirical	$\frac{Z}{(i+B)^C}$
1	11	11.42
2	9	9.14
3	9	7.72
4	7	6.74
5	6	6.01
6	6	5.45
7	5	4.99
8	5	4.62
9	4	4.31
⋮	⋮	⋮
15	4	3.15
16	3	3.02
⋮	⋮	⋮
21	3	2.53
22	2	2.46
⋮	⋮	⋮
39	2	1.67
40	1	1.65
⋮	⋮	⋮
124	1	0.75
$Z = 225 \cdot 0.103 = 23.175$ $B = 1.7139$ $C = 0.709$		

where V_F, V_F^* are taken from Tab. 3 and $\hat{V}$ is the average of the V_F ($\hat{V} = 124/11 = 11.27$). The test criterion (14) can be applied to test the fit of an arbitrary curve to the empirical data (cf. Yamane 1964: 803). Normally a fit is considered as "good", when **Krit** > 0.9.

For the example above (14) yields

Table 3
FD of the word forms in Goethe's "Erlkönig"

Frequency F	Number of occurrences	
	empirical V_F	$V_F^* = \frac{\hat{Z}}{(F+\hat{B})^{\hat{C}}}$
1	85	84.76
2	18	18.76
3	6	7.47
4	7	3.85
5	2	2.29
6	2	1.49
7	1	1.04
8	0	0.76
9	2	0.57
10	0	0.45
11	1	0.36
$\sum$	124	121.8

$$\textbf{Krit} = 1 - \frac{16.299}{6250,2} = 0.997,$$

indicating that the fit is very good.

4. A modification of the FD-model

Finally we consider a simple modification of the FD-model given by (13). We substitute $\hat{Z}$ for a standardizing constant K, i.e. we take the approximation

$$V_F \approx V_F^{(1)} := \frac{K}{\left(F+\hat{B}\right)^{\hat{C}}}, \tag{15}$$

where $\hat{B}$, $\hat{C}$ are given as above and

$$K = V\left(\sum_{F=1}^{11}\frac{1}{\left(F+\hat{B}\right)^{\hat{C}}}\right)^{-1}$$

Table 4
FD of the word forms in the "Erlkönig" of Goethe

Frequency F	Number of occurrences		
	empirical V_F	$V_F^{(1)}$	$V_F^{(2)}$
1	85	86.29	81.12
2	18	19.10	20.90
3	6	7.60	8.78
4	7	3.92	4.63
5	2	2.33	2.79
6	2	1.52	1.84
7	1	1.06	1.28
8	0	0.77	0.94
9	2	0.58	0.71
10	0	0.46	0.56
11	1	0.37	0.44

The brackets show how the classes are pooled in the chi-square test.

(cf. section 1). For our example we obtain

$$V = 124, \quad K = 124\left(\frac{121.8}{\hat{Z}}\right)^{-1} = 120.8$$

(cf. Tab. 3). The theoretical values $V_F^{(1)}$, obtained by (15) are listed in the third column of Tab. 4. The chi-square test yields $\chi^2 = 5.21$. The number of degrees of freedom is 5, since the theoretical values have to be pooled to classes > 1 (cf. Altmann 1988: 74) This yields the corresponding probability 0.39. The fourth column in Tab. 4 shows the optimal fit

$$V_F^{(2)} = \frac{K'}{(F+B')^{C'}},$$

obtained by the Nelder-Mead procedure ($K' = 168.48$, $B' = 0.3484$, $C' = 2.4449$).

In this case the chi-square test yields $\chi^2 = 3.05$ with 6 degrees of freedom and the corresponding probability $P = 0.80$. The comparison between $V_F^{(1)}$ and $V_F^{(2)}$ shows that the fit of $V_F^{(2)}$ is better altogether, but because of the choice of $\hat{B}$ (section 3) the prediction by $V_F^{(2)}$ is better for small F.

Literature

Altmann, G. (1988). *Wiederholungen in Texten.* Quantitative Linguistics, Vol. 36. Bochum: Brockmeyer.

Arapov, M.V. (1988). *Kvantitativnaja lingvistika* (Quantitative Linguistics). Moskva: Nauka.

Boroda, M.G. (1982). "Häufigkeitsstrukturen musikalischer Texte." In Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.Š. (1982) Sprache, Text, Kunst. Quantitative Analysen. Bochum: Brockmeyer. 231-262.

Boroda, M.G. (1990). "The Organization of Repetitions in the Musical Composition: Towards a Quantitative-Systemic Approach. In *Musikometrika* 2. Bochum: Brockmeyer. 53-106.

Boroda, M.G., Zörnig, P. (1990). "Zipf-Mandelbrot's Law in a Coherent Text: Towards the Problem of Validity." In *Glottometrika* 12. Bochum: Brockmeyer.

Haight, F.A. (1969). "Two Probability Distributions connected with Zipf's Rank-Size Conjecture." *Zastosowania Matematyki Applicationes Mathematicae*, Vol. X.

Hajtun, S.D. (1983). *Naukometrija-sostojanie i perspektivy* (Sciencemetrics – State-of-the-Art and Perspectives). Moskva: Nauka.

Herdan, G. (1966). *The Advanced Theory of Language as Choice and Chance.* Berlin, Heidelberg, New York: Springer Verlag.

Krallmann, D. (1966). *Statistische Methoden in der stilistischen Textanalyse.* Bonn.

Krylov, Ju.K. (1987). "Stacionarnaja model porožđenija svjaznogo teksta" (A Stationary Model of Generation of a Coherent Text). *Acta et Commentationes Universitatis Tartuensis* 774, 81-102.

Orlov, Ju.K. (1982). "Ein Model der Häufigkeitsstruktur des Vokabulars." In *Studies on Zipf's Law.* Quantitative Linguistics, Vol. 16. Bochum: Brockmeyer.

Tuldava, J. (1987). *Problemy i methody kvantitativno-sistemnogo izučenija leksiki.* (Problems and Methods of Quantitative-Systemic Study of Lexis). Tallinn: Valgus.

Yamane, T. (1964). *Statistics: An Introductory Analysis.* New York: Harper & Row.

Rieger, Burghard (ed.)
Glottometrika 13
Bochum: Brockmeyer 1992, 219-229

Relationships in the Length, Age and Frequency of Classical Japanese Words

Tatsuo Miyajima

1. Frequency and length of words

The exercise of a little common sense will enable us to realize that there is a close relationship between the frequency and the length of words, and that short words are used more often. This was already proven by Zipf on the basis of data in different languages (Zipf 1935). The purpose of this report is to confirm a similar result for the words used in Japanese classical works.

First I must explain the corpus (the classical works) and the frequency table used here. "Koten taishō goihyō" (Miyajima 1971) is a frequency table of words used in the 14 most representative Japanese classics listed below.

	different words	total numer of words
Man-yōshū (8c. anthology)	6505	50070
Taketori Monogatari (9c. romantic tale)	1312	5124
Ise Monogatari (9c. romantic tale)	1692	6931
Kokin Wakashū (10c. anthology)	1994	10015
Tosa Nikki (10c. travel diary)	984	3496
Gosen Wakashū (10c. anthology)	1923	11955
Kagerō Nikki (10c. diary)	3598	22398
Makura no Sōshi (11c. essay)	5246	32905
Genji Monogatari (11c. The tale of Genji)	11421	207792

	different words	total numer of words
Murasaki-shikibu Nikki (11c. diary)	2468	8737
Sarashina Nikki (11c. travel diary)	1950	7243
Ōkagami (11c. historical tale)	4819	29212
Hōjōki (13c. essay)	1148	2527
Tsurezuregusa (14c. essay)	4240	17112
Total	23877	414417

The length of words was calculated from the number of kana letters. Kana are a Japanese syllabary and a kana basically represents a syllable (strictly speaking, a mora). Sometimes two kana represent a syllable, but such a discrepancy between letters and sounds occurred less frequently in old Japanese than nowadays. In the following I use for convenience the expression 'a word with X syllables' instead of a more exact one like 'a word represented with X kana letters'.

Japanese texts are written without a space between words. The segmentation of texts into words, and accordingly the result of word-counts, largely depends on the theories of scholars. For example, the number of words in the Tale of Genji is estimated 200,000 or 400,000 depending on the point of view. I do not enter into detail here and refer you for the rule of segmentation to "Koten taishō goihyō". One thing I want to mention is that 'joshi' and 'jodoshi' which are sometimes regarded as words and translated as 'particle' and 'auxiliary verb' are treated here as suffixes and were not counted.

The longest 'words' among the corpus are the following indecipherable phrases:

"mumanokituriyaukituninowokanakakuboreirikurentou" (Tsurezuregusa, 26 syllables)
"tunokamenadonitatetekufumonomatukaikake" (Makura no Soshi, 20 syllables)

Other long ones are year numbers, official titles and sutra titles.

The following table shows the total results of the 14 works.

[Table 1]

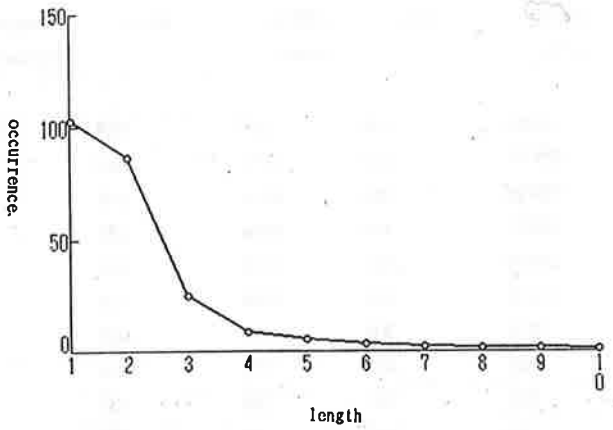
length of words	running words	(ratio)	different words	(ratio)	average occurrences
1	22391	.054	218	.009	102.71
2	175597	.423	2031	.085	86.46
3	108726	.264	4330	.181	25.34
4	60872	.146	6846	.287	8.89
5	29919	.072	5128	.215	5.83
6	11813	.028	3100	.130	3.81
7	3819	.009	1496	.063	2.55
8	920	.002	453	.019	2.03
9	304	.001	168	.007	1.81
10	81	.000	63	.003	1.29
11	46	.000	18	.001	2.56
12	11	.000	9	.000	1.22
13	5	.000	5	.000	1.00
14	4	.000	4	.000	1.00
15	1	.000	1	.000	1.00
16	1	.000	1	.000	1.00
17	4	.000	3	.000	1.33
18	1	.000	1	.000	1.00
20	1	.000	1	.000	1.00
26	1	.000	1	.000	1.00

average number of syllables 2.90.

It is clear that the longer a word is, the less it is used.

This research was carried out on the length of the entry forms of words, but the inflected forms in texts show a similar tendency (Ishii 1990).

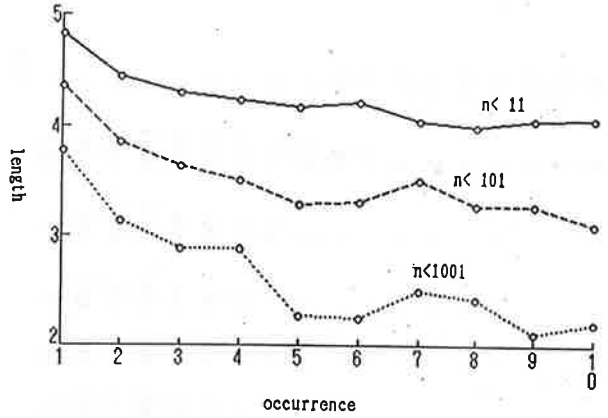
The high frequency of short words must be a universal tendency observed in many languages. But the number of different syllables is limited in Japanese because most of them are open, which leads to an abundance of homonyms among short words. So, unlike English or Chinese, polysyllabic words are preferred to monosyllabic ones.



If we take the opposite point of view and observe the length of words based on their frequency, we get the same results: the more frequently a word occurs, the shorter it is. We give below the total results but the same conclusion is arrived at in every individual work.

[Table 2]

occurence	length				
n= 1	4.83	n<11	4.36	n<101	3.77
n= 2	4.45	10<n<21	3.85	100<n<201	3.13
n= 3	4.30	20<n231	3.62	200<n<301	2.88
n= 4	4.23	30<n<41	3.49	300<n<401	2.88
n= 5	4.17	40<n<51	3.27	400<n<501	2.28
n= 6	4.21	50<n<61	3.30	500<n<601	2.25
n= 7	4.04	60<n<71	3.49	600<n<701	2.50
n= 8	3.99	70<n<81	3.26	700<n<801	2.43
n=9	4.05	80<n<91	3.26	800<n<901	2.11
n=10	4.06	90<n<101	3.09	900<n<1001	2.21
				1000<n	2.16



As for the number of different words, the most frequent ones are those with 4 syllables and then come those with 5, 3, 6, 2, 7 ... syllables in this order. Monosyllabic words are low on the scale. The same is true of modern japanese. In the next table, we arranged the length of words according to their number and compared them to data in modern Japanese. The latter were taken from Hayashi (1957) who counted the words in NHK's "Nihongo akusento jiten" (The Japan Broadcasting Corporation's "Japanese Accent Dictionary"). The ratio is given in parentheses.

[Table 3]

order	old Japanese		modern Japanese	
	length	ratio	length	ratio
1	4	6846(.287)	4	(.388)
2	5	5128(.215)	3	(.227)
3	3	4330(.181)	5	(.177)
4	6	3100(.130)	6	(.110)
5	2	2031(.085)	2	(.048)
6	7	1496(.063)	7	(.033)
7	8	453(.019)	8	(.012)
8	1	218(.009)	1	(.003)
9	9	168(.007)	9	(.002)
10	10	63(.003)	10	(.001)

[Table 4]

length	number of works														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	average
1	92	29	11	17	13	8	5	7	5	4	5	7	4	11	4.06
2	861	240	145	121	101	90	70	63	62	39	55	50	42	92	4.08
3	2376	570	325	250	202	128	117	63	64	67	59	45	31	33	2.75
4	4383	1020	507	339	208	140	97	57	37	23	21	8	5	-	1.94
5	3646	745	344	171	93	61	31	13	13	3	4	1	3	-	1.61
6	2512	370	115	49	25	13	4	3	5	2	2	-	-	-	1.33
7	1325	112	32	14	8	2	2	-	-	1	-	-	-	-	1.19
8	414	27	9	3	-	-	-	-	-	-	-	-	-	-	1.12
9	159	6	3	-	-	-	-	-	-	-	-	-	-	-	1.07
10	62	1	-	-	-	-	-	-	-	-	-	-	-	-	1.02
11	18	-	-	-	-	-	-	-	-	-	-	-	-	-	1.00
12	9	-	-	-	-	-	-	-	-	-	-	-	-	-	1.00
13	5	-	-	-	-	-	-	-	-	-	-	-	-	-	1.00
14	4	-	-	-	-	-	-	-	-	-	-	-	-	-	1.00
15	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1.00
16	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1.00
17	3	-	-	-	-	-	-	-	-	-	-	-	-	-	1.00
18	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1.00
20	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1.00
26	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1.00
total	15874	3120	1491	965	649	442	326	206	186	139	146	111	85	137	23877
average occurrences	1.9	5.2	10.0	16.4	25.3	38.0	57.2	66.1	120.6	148.4	176.5	204.6	391.3	1087.4	17.4

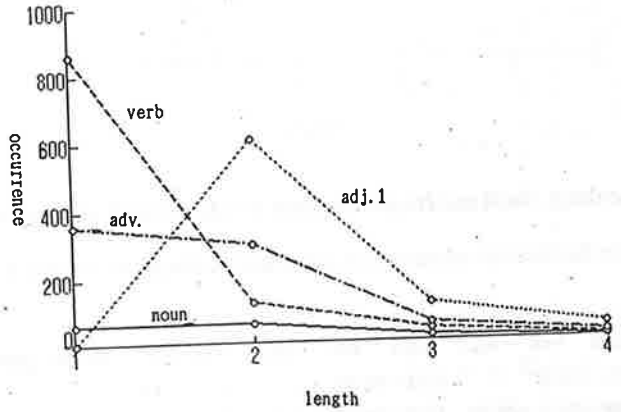
Generally speaking, short and frequently used words are used in many works.

Next, we give the average occurrences according to the parts of speech.

[Table 5]

length	noun	verb	adj.1	adj.2	adv.	adn.	con.	int.	epithet	phrase
1	59.24	836.20	-	31.00	356.33	-	-	1.60	-	-
2	59.48	122.23	608.79	72.87	300.02	843.11	5.67	35.24	-	-
3	15.25	38.02	117.40	36.32	54.08	17.50	57.83	4.86	-	5.36
4	6.93	8.38	44.36	15.72	19.89	4.33	11.25	1.57	2.57	1.00
5	5.17	5.20	13.27	7.09	10.39	-	-	-	10.27	6.52
6	3.60	3.16	15.53	2.64	6.14	-	3.00	1.00	-	1.00
7	2.74	2.06	6.68	1.79	2.33	-	-	-	-	2.63
8	2.49	1.48	2.44	1.50	12.50	-	-	-	-	1.00
9	2.34	1.31	1.47	1.00	-	-	-	-	-	-
10	1.37	1.20	1.00	1.00	-	-	-	-	-	1.00
11	2.87	1.00	-	-	-	-	-	-	-	-
12	1.17	2.00	-	1.00	-	-	-	-	-	1.00
13	1.00	1.00	-	-	-	-	-	-	-	-
14	1.00	-	-	-	-	-	-	-	-	1.00
15	1.00	-	-	-	-	-	-	-	-	-

length	noun	verb	adj.1	adj.2	adv.	adn.	con.	int.	epithet	phrase
16	1.00	-	-	-	-	-	-	-	-	-
17	1.33	-	-	-	-	-	-	-	-	-
18	1.00	-	-	-	-	-	-	-	-	-
20	-	-	-	-	-	-	-	-	-	1.00
26	-	-	-	-	-	-	-	-	-	1.00
(average length)										
	2181	2.97	3.37	3.36	2.50	2.01	3.09	2.08	4.97	5.05



Adjectives 1 and adjectives 2 have similar meanings and functions but different morphological forms. Adnominals have formal meaning and always modify nouns. Entry forms of adjectives 1 and adnominals have more than one syllable. Epithets are idiomatic modifiers of certain nouns and are used in poems.

Monosyllabic verbs have high frequencies because some basic verbs with formal usage like "su(do), fu(pass), ku(come), u(get)" are used very often. Many of the monosyllabic nouns, for example "yo(world), mi(body), me(eye), hi(day), te(hand)", are also basic, but their meanings seem to be more substantial than verbs.

Old Japanese adopted many borrowings from Chinese. (They were, however, rarely used in poems.) The following table shows the average occurrences of words classified by length and origin.

[Table 6]

length	native Japanese	Chinese borrowings	hybrids
1	147.42	10.15	-
2	104.68	16.00	6.73
3	29.66	7.65	9.86
4	9.57	4.88	7.17
5	6.02	5.20	3.26
6	3.78	4.84	2.73
7	2.38	2.52	4.18
8	1.97	2.18	2.09
9	1.32	2.64	1.89
10	1.16	1.52	1.00
>10	1.10	1.83	2.00
(average length)			
	2.85	3.47	4.59

Hybrids have, by definition, two or more components and cannot be monosyllabic. Their average length is longer than native Japanese words and borrowings of Chinese origin.

Native Japanese words rank from the first to the tenth in the reverse order of length. But among the Chinese borrowings first come disyllabic words, then monosyllabic, and then trisyllabic. The reason why monosyllabic borrowings do not occur as often as native ones is probably that nearly all of them are nouns. While many of the monosyllabic words with high frequency are verbs, borrowings are generally nouns. It is necessary to convert them into hybrids by adding "su(do)" in order to use them as verbs.

2. Frequency and age of words

One can easily suppose that words used since olden times are more basic and occur more frequently than others. We now confirm this fact on the basis of the data in "Koten taishō goihyō".

We divided the words in "Tsurezuregusa" into the following three groups:

Old: those which appeared in Man-yōshu
Middle: those which do not appear in Man-yōshū but in Genji Monogatari
New: those which do not appear in either of them

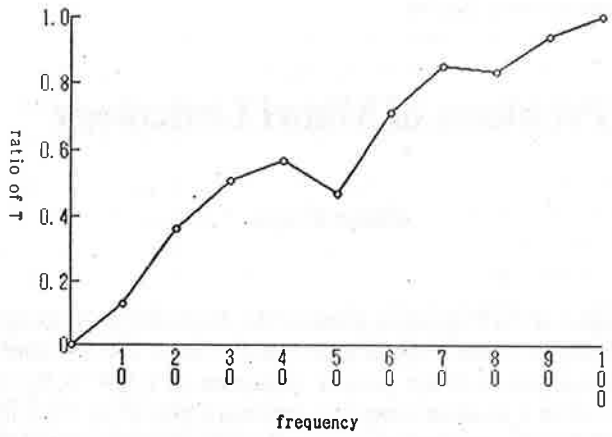
Average frequency and length in Tsurezuregusa of each group are given below:

	running words	different words	average occurence	average length
Old:	10729	1050	10.22	2.76
Middle:	3657	1272	2.88	3.79
New:	2726	1918	1.42	4.23

Clearly older words are shorter and used more frequently. In other words, those words with high frequency would problably remain in later times. Let us support this argument by examining the words in Man-yōshū . We classified them according to frequency and checked their appearance in Tsurezuregusa. *T* indicates the number of words used in Surezuregusa and *NT* the number of those words which did not occur in it.

[Table 7]

	T	NT	ratio of T
11>n	652	5079	.127
21>n>10	131	233	.360
31>n>20	62	60	.508
41>n>30	37	28	.569
51>n>40	23	26	.469
61>n>50	22	9	.710
71>n>60	17	3	.850
81>n>70	20	4	.833
91>n>80	16	1	.941
101>n>90	8	-	1.000
n>100	62	12	.838
average frequency	21.85	3.69	
average length	2.76	3.93	



Many of the words which remained in the time of Tsurezuregusa had shorter forms and high frequency in Man-yoshu.

References

Hayashi, Oki (1957), *Goi* [Vocabulary]. Koza Gendai Kokugogaku II.

Ishii, Hisao (1990), *Zasshi ni okeru go nagasa* [The length of words in a magazine]. Keiryo Kokugogaku.

Miyajima, Tatsuo (1971), *Koten taishō goihyō* [A frequency table of words in classical Japanese works]. In 1989 this was published in a floppy disk for use in personal computers.

Zipf, G. (1935), *The psycho-biology of language.*

Rieger, Burghard (ed.):
Glottometrika 13
Bochum: Brockmeyer 1992, 230-240

Problems of Maori Lexicology

Viktor Krupa

Maori, a member of the Polynesian group of the Austronesian language family, displays a variety of extreme typological characteristics that are shared by all Polynesian languages, with the possible exception of a few of the so-called Outliers. The latter, a result of secondary westward migrations from Polynesia proper into Melanesia, are spoken in several small islands scattered along the main islands and exposed to large-scale interference with the local non-Polynesian languages.

The state of the Maori lexicon cannot be satisfactorily characterized without a brief recourse to social and political background of the language. The Maoris have inhabited New-Zealand since about thousand years and their language has been evolving in a complete isolation from other languages right until the end of the 18th century. The ever expanding contacts with the British immigrants and their civilization have resulted in a gradual and obviously irreversible retreat of the Maori language which may ultimately lead to its extinction in most functions with the most obvious exception of the ceremonial function.

The only available dictionary has been used as a source for the present work, i.e., Herbert W. Williams' book titled "A Dictionary of the Maori Language" published in many editions (6th edition printed in 1957 and the 7th edition published in 1971 and reprinted in in 1988). The dictionary contains some 15,000 entries. However, only original Maori words are listed by the compiler and borrowings (mainly from English) are given in Appendix, with the remark that it comprises some of the more important words adopted from non-Polynesian sources. This solution seems to be inadequate at the first glance but it is very hard to decide, at least nowadays, where is the boundary between assimilated borrowings and massive lexical interference in the speech of modern Maoris who are fully

bilingual in English and Maori - if they are fully competent speakers of the Maori language at all.

The size of Maori vocabulary is moderate when compared to other languages included in the project, which is due the historical circumstances (absence of advanced technology, absence of writing, isolation, etc.) in pre-European era and to a gradual abandonment of the Maori language in favour of English.

The typological characteristics of the Maori language are thought to be relevant for the reaction of the Maori population to the practical value of their language vis-à-vis English in the critical contact situation. This reaction has obviously decided the fate of the Maori language in the future. The most obvious structural consequences of the retreat of the language are a continuing simplification of its grammatical categories and an analogous development in its vocabulary.

Information on Maori words is provided mainly in columns 1, 4, 5, 6, 7, 9, 14 of the questionnaire of the project "Language synergetics".

A. The discussion of the typological peculiarities will be opened with phonology because this is of immediate relevance to the transcription used in Column 1. The latter differs from the official orthography of Maori only in two respects: (1) the digraph *ng* has been replaced by *g*, (2) the digraph *wh* has been replaced by f.

The phonological inventory of Maori is extremely meagre, comprising ten consonants (p, t, k, m, n, ng, w, r, h, wh) and five vowels (a, e, i, o, u). All five vowels have their long pendants (written as either aa, ee, ii, oo, uu or $\bar{a}$, $\bar{e}$, $\bar{i}$, $\bar{o}$, $\bar{u}$). The digraph *ng* denotes a velar nasal sound close to English *ng* in, e.g., going, and the phonetic realization of *wh* varied a good deal geographically. Lately its pronunciation as English voiceless labiodental *f* has prevailed over the more archaic, likewise voiceless, bilabial pronunciation.

B. The scarcity of Maori phonology is not restricted to its inventory of phonemes but is likewise present in its combinatorics, in phonotactics. There are very few types of syllables - V, VV, CV, CVV - which may be schematized as (C)V(V). The syllabification rules have been summed up by P.W. Hohepa; according to him, syllable boundaries occur at word space, before each consonant, after identical vowels in close transition, before identical vowels preceded by another vowel in close transition, and after every second vowel in a non-identical close transition sequence (Hohepa 1967: 9). The so-called long vowels are regarded as sequences of two identical vowels, which makes the treatment of

Maori syllabification and stress much simpler (cf. Biggs 1961, Hohepa 1967). The long vowels as sequences of two identical vowels have appeared as a consequence of the loss of a consonant attested for Proto-Austronesian, e.g., Maori *too* "stem" from PAN *tebu, *raa* "sail" from PAN *layaR, *hoo* "give" from PAN *beRay. Other long vowels have appeared on the joint of two morphemes, e.g. *pana* "drive away + *-a* "passive marker" = *panaa* "be driven away". In accordance with this view, long vowels have been suggested by B. Biggs to be written as *aa*, *ee*, *ii*, *oo*, *uu*.

Stress in Maori is positional and therefore cannot distinguish meaning. Major stress falls on the first syllable containing a sequence of identical vowels; if there is none, it falls on the first syllable with a non-identical sequence of vowels, and, finally, if there is no such sequence, it falls on the first syllable. The final unstressed syllable tends to be devoiced (Hohepa 1967: 10).

C. The phonotactic rules do not admit any consonantal clusters at all. In the speech flow, all syllables end in a vowel and each consonant must be followed by a vowel. However, a morphemic analysis may lead to the discovery of a few consonantal morphemes such as, e.g., *t-*, *m-*, and *n-* in possessive pronouns (*taana* versus *toona* "his, hers", *maana* versus *moona* "for him, for her", *naana* versus *noona* "by him, by her".

Unlike consonants, vowels may enter into a variety of clusters, some of them four, five, or even six vowels long, e.g. *taaua* "we (dual, inclusive)", *tuuaaumu* "spell"; *aeaeaa* "panting".

D. Heavy phonotactic restrictions limit the number of syllables. There are only 55 syllables structured (C)V in Maori, including 4 syllables occurring only in loanwords from English that are of a relatively late date (*wo*, *wu*, *who*, *whu*). Besides, there are syllables of the type (C)VV; diphthongs *ai*, *au*, *ei*, *eu*, *ou* occur in the vocalic part of this syllable type, as well as sequences of identical vowels (Krupa 1968: 21-28).

The absence of consonantal clusters and the obligatory presence of a vowel after each consonant have far-reaching consequences for the morpheme structure. Another limitation is of a statistical nature: most morphemes contain two vowels. Shorter morphemes are rare but very frequent while morphemes containing more than two vowels, are rare and infrequent. The stock of Maori morphemes (to be precise, of morphemic forms) follows a few basic patterns. The number of vowels is the chief classificatory criterion. According to the latter, the morphemic forms

are subdivided into mono-vocalic, bi-vocalic, and poly-vocalic. An additional criterion to be introduced is the number and position of the consonants within the morphemic forms. Now the stock of morphemic forms in Maori may be divided into mono-vocalic, bi-vocalic, and poly-vocalic and all of them may in their turn be subdivided into full (CV, CVCV, CVCVCV, ...) as well as reduced initially (V, VCV, VCVCV, ...), medially (CVV, CVVCV, ...) or completely (VV, VVV, ...).

Very strong phonotactic restrictions applied to a likewise restricted phonological inventory lead to a far-reaching consequence - the total morphemic stock of Maori is quite modest. It consists of a few tens of mono-vocalic morphemic forms reserved for grammatical functions (highly homonymous), of several hundreds of bi-vocalic forms (the upper theoretical limit is 3,025 but due to further combinatorial restrictions only 1,258 such forms are really admissible) and of a fairly small category of poly-vocalic forms; the morphemic forms of the two latter types function as roots of autosemantic words.

The approximately 1,300 bi-vocalic morphemic forms are the material of which the vast bulk of the vocabulary of Maori is built up. This inevitably implies a fairly high degree of homonymy. Upon the basis of a sample of 100 bi-vocalic forms selected from the dictionary, it has been found out that they manifest 227 homonymous semes. This means that the index of homonymy $I_H = 227/100 = 2.27$. An index of polysemy may be calculated in a similar way. Since the 227 bi-vocalic morphemes display 593 various meanings, the index of polysemy $I_P = 2.61$. We could use the homonymy index to estimate the number of all bi-vocalic morphemes in Maori, which could be calculated as $I_H \times 1,258 = 2855$, i.e., less than 3,000 morphemes.

Poly-vocalic morphemes are adjusted loanwords from English, e.g., *taakuta* "doctor", *pirimia* "prime minister", but quite a few of them have been inherited from proto-language, e.g., *whenua* "land, country", *aroha* "love, pity".

E. Column 4 gives the word class characteristics of each word. Most grammatical functions in Maori are carried by special, usually proclitic markers called particles of which there are two basic classes, nominal and verbal. This explains scarcity of inflective affixes in Maori. Neither are derivative affixes very common.

The scarcity of inflective markers makes the definition of word classes in Maori quite complicated and one has to rely more on the distribution than on the internal word structure. Some authors were inclined to define word classes in Maori as

syntactic categories, which amounts to saying that in principle any word is capable of functioning either in a nominal or in a verbal phrase. However, this is not hundred per cent true although conversion is a common phenomenon. The following word classes are distinguished here, i.e., interjections (Int), grammatical particles (Pt_g), modal particles (Pt_m), nouns (N), locatives (L), adjectives (A), statives (St), verbs (V), adverbs (Adv), pronouns (Pr), numerals (Num), and conjunctions (Conj).

The interjections, unlike the remaining word classes, do not enter syntactic constructions and remain outside sentence structure. Aside from interjections, all words can be divided into two large groups of classes, i.e., into particles and full words (sometimes called bases).

Particles are enclitic, usually monosyllabic words that modify the grammatical or the lexical meaning of the full words with which they combine to create phrases. There are two basic classes of particles - grammatical and modifying particles. Grammatical particles are typically prepositive (in relation to the phrase nucleus) and represent two mutually exclusive sets, i.e., verbal (or processual) particles marking mood/aspect/tense and nominal particles marking categories of definiteness, case, spatial relations, etc. On the other hand, the modifying (or complementary) particles may and do occur in both nominal and verbal phrases, being at the same time typically postpositive and bi-vocalic.

The class of nouns is open and includes words compatible with the (determinative) nominal particles and incompatible with the verbal particles. Their typical syntactic functions are subject, object and nominal predicate.

Locatives are a small closed class of words incompatible with the verbal particles and with the nominal determinative particles, being compatible only with prepositions. Their typical syntactic function is location of the action (or event) or of an object referred to by the sentence.

The open class of adjectives comprises words compatible both with nominal and verbal particles but incompatible with the passive suffix. Their prototypical syntactic functions are predicate (compatibility with *he* "indefinite article") and attribute (postpositive to the nominal nucleus).

A peculiarity of the Polynesian languages is the small closed class of statives (sometimes labelled participles) that are compatible with the verbal particles and incompatible with the nominal particles. Because of their static nature, they are incompatible with the passive marker as well. They usually occur in the function

of the verbal predicate. Their agent is marked with the oblique case marker *i* while their syntactic subject is the target of the action referred to by a stative.

The largest class is that of verbs, sometimes also labelled universals because most of them may function as syntactic nouns without any further modification. The class of verbs is notable for its compatibility with both the verbal and the nominal particles, as well as with the passive suffix and in this sense their distribution may be said to be universal. In practice, however, some of the verbs usually occur only in nominal constructions while others are preferably employed in verbal phrases.

Adverbs represent a fairly small class of words that modify the lexical meaning of the verbal phrase nuclei, many of them functioning as intensifiers (Int in Column 7).

Pronouns have been set up as a special class in view of their semantics. As far as their morphology and distribution is concerned, they are a rather heterogeneous lot. The following prefixes have been found to occur with some sub-classes of pronouns: *pee-* "comparative" (= like), *tee-* "determinative sing.", *ee-* "determinative plur.", possessive prefixes of the series *t-*, *n-*, *m-*, *o-* (*taa-*, *too-*, *naa-*, *noo-*, etc.). Dual and plural of the personal pronouns are marked by suffixes *-(r)ua* and *-tou* respectively.

The class of numerals is a closed set of words defined semantically, just as pronouns, and syntactically reminiscent of adjectives. They are compatible with the prefixes *tua-* "ordinal", *taki-* "distributive", *toko-* "human numerative", and *ee-* "non-human numerative".

Conjunctions are another heterogeneous and nascent class of words including very few primary conjunctions such as *aa* "and", *engari* "but", and several for the most part nominal expressions used to link clauses.

F. Column 5 is relevant for the class of verbs (universals) that are compatible with the passive suffix. This suffix has 12 variants in Maori, i.e. 1. *-a*, 2. *-ia*, 3. *-ina*, 4. *-na*, 5. *-kia*, 6. *-mia*, 7. *-ngia*, 8. *-ria*, 9. *-tia*, 10. *-whia*, 11. *-hia*, 12. *-nga*. There is also the nominalizer *-((C)a)nga* available in seven variants, i.e., 1. *-nga*, 2. *-anga*, 3. *-hanga*, 4. *-kanga*, 5. *-manga*, 6. *-ranga*, 7. *-tanga*.

The verbs may be subdivided into twelve respective formal classes according to their preferences for a particular variant. Several verbs may freely combine with more than one variant. On the whole it holds that in spite of a certain degree of

free variation and of a low productivity of some of the variants (e.g., *-mia*, *-ngia*, *-whia*), a set of statistical rules may be formulated to account for more than 80 % of the instances: 1. Bi-vocalic verbs ending in -a prefer the variants *-a*, *-ia*, *-ina*, 2. Those ending in *-e* take *-a*, 3. Those ending in *-i* usually prefer *-a*, 4. Those ending in *-u* in many cases combine with *-a*, 5. Those ending in *-ki* prefer *-na*, 6. Poly-vocalic verbs usually prefer the variant *-tia* and to a much lesser degree the variant *-a*.

The distribution of the nominalizer variants is on the whole parallel to that of the passivizer.

The profusion of the passive suffix variants is obviously due to the reinterpretation of the verbal root in the proto-language for which the structure $C_1V(C)C_2VC_3$ was typical and C_3 has (at least partly) been preserved only in the passive voice and reinterpreted as its initial consonant. There is a tendency to eliminate many of the variants, especially in favour of two variants, i.e., *-tia* and *-a*. However, the semantics of the passive voice is also in flux; many Polynesian languages display ergative sentence structure and in Maori itself the passive constructions are incomparably more frequent than their active pendants, which is very peculiar.

The list of inflective affixes includes also the only highly productive causative prefix *whaka-*. All words containing this prefix are verbs, except those that contain also the nominalizer.

The two affixes, causative and nominalizer cannot be unambiguously classed as either inflective or derivative because they merge both types of functions, which is typical of many Austronesian languages.

G. The morphological status is given in Column 6 as either stem (S), derivate (D), compound (C), reduplicate (R), uncertain status (U), and unknown (N).

Pure derivative affixes may be regarded as non-productive relics. They are (1) prefixal stem-formants *aa- # a-*, *haa- # ha*, *hii- # hi*, *hoo- # ho-*, *huu- # hu-*, *ii- # i-*, *kaa- # ka-*, *kii- # ki-*, *koo- # ko-*, *kuu- # ku-*, *maa- # ma-*, *moo- # mo-*, *ngaa- # nga-*, *oo- # o-*, *paa- # pa-*, *pii- # pi-*, *poo- # po-*, *puu- # pu-*, *raa- # ra-*, *rii- # ri-*, *roo- # ro-*, *ruu- # ru-*, *taa- # ta-*, *tii- # ti-*, *too- # to-*, *tuu- # tu-*, *uu- # u-*, *whaa- # wha-*, *whee- # whe-*, *whii- # whi-*. A word as a rule contains only one of these prefixes but there is a tendency to merge them with the stem. Their meaning is rather opaque and could be described as static or qualitative; (2) suffixal stem-formants *-aa # -a* and *-hi # -ki # -mi # -i*. The suffix *-aa # -a* has also the

meaning of a state or quality while the suffixes *-hi # -ki # -mi # -i* have a general meaning of action or state.

Due to the non-productivity of derivative affixes, it is reduplication and composition that play key role in the extension of vocabulary. There are two basic types of reduplication in Maori - partial and complete reduplication. Reduplication in Maori (unlike, e.g., Indonesian) is in principle a matter of the morpheme, not of the word. It is always the root that is reduplicated and in case of a compound only one of the two (or more) root morphemes may be reduplicated, e.g., *poohue* "a climbing plant" = *poopoohue* = *poohuehue* "dtto". Thus the occurrence of a reduplicated form in the text is very helpful for morphemic analysis without being completely reliable. In quite a few instances the structure of a word is reinterpreted in the spirit of folk etymology. For example, *taina* "younger sibling", originally consisting of the root *tai-* and the suffix *-na* has been reinterpreted as *ta- + -ina* (i.e., quasi-prefix + quasi-root) and reduplicated as *taainaina* "younger siblings".

Partial reduplication may be characterized as the repetition of the first syllable of a root, e.g., *patu* "to strike, beat" - *papatu* "to beat one another, clash". Partial reduplication marks, according to B. Biggs, a single terminal action, e.g., *papaki* "slap or clap once", diminished intensity, e.g., *papango* "somewhat black", and in a few cases plural, e.g. *raakau roroa* "tall trees" (Biggs 1969: 107); however, other sources mention partial reduplication as a marker of reciprocal action, cf. *papatu* "to beat one another" and frequency, e.g. *kakai* "to eat frequently" (Williams 1957: 86).

Complete reduplication amounts to the repetition of the whole root, e.g., *mate* "to die" - *matemate* "to die in numbers". Usually, complete reduplication marks frequency of action, e.g., *haaereere* "to stroll, wander about", extended duration, e.g., *kaukau* "to bathe" (i.e., to swim for some time), or plurality, cf. *matemate*. These semantic nuances seem to correlate, at least partly, with the word class characteristics. Thus partial reduplication marking plural is attested for adjectives, just as complete reduplication marking diminished intensity; complete reduplication with a plural function is typical for verbs and the same is true of frequentative and durative meanings.

Reduplication serves as a means of marking some abstract meanings or grammatical functions and is also employed as a device of vocabulary extension, and it is very difficult to distinguish its inflective and derivative functions. In Column 7, most of the reduplicated forms are indexed as Int+ or Int-, i.e., increase or

decrease in the intensity of meaning, Rec, i.e., reciprocal action, Freq, i.e., frequentative action, and Sim, i.e., similitiveness.

Compounds usually consist of two autosemantic root morphemes where signifiant is following signifié and are attested for most of the word classes, including verbs.

H. Some functional properties of nouns, verbs, pronouns, and particles are specified in Column 7. In addition to categorial types of reduplication, grammatical particles are coded as either nominal (n) and verbal (v). Within the class of nominal particles, the subclass of determinatives is coded as Det; all markers sensitive to number are specified as singular (Sing) or plural (Pl) and the category of possession (Poss) covering alienable and inalienable property is coded as Al and Inal. Pronouns are subclassified into personal (Pers), possessive (Poss), interrogative (Qu), and demonstrative (Dem) and characterized as Al and Inal as well as Sing and Pl where necessary.

Verbal particles may be specified as imperfective (Imp), perfective (Perf), future (Fut), inceptive (Inc), and anaphoric (An).

The grammatical structure of Maori requires, however, further specification. The nominal grammatical particles are grouped into two subclasses, i.e., determinative particles (*te* "definite article sing.", ngaa "definite article plural", *a # aa* "personal article", *he* "indefinite article singular" (in plural replaced by *eetahi* "several, some"). The nouns themselves are not inflected for number, with the exception of a handful of nouns, such as *tane* "man" - *taane* "men", *wahine* "woman" - *waahine* "women", *tangata* "human being" - *taangata* "people", etc. The basic form may in principle refer to more than one object. The plural article *ngaa* is only employed when it is necessary to stress the plurality of objects. The determinative particles take up the slot immediately before the head of the nominal phrase. They may be preceded by another class of nominal particles, i.e., prepositions, e.g. *ki* "to, toward", *me* "with", *e* "by", *i* " accusative, with, from, agentive with the statives, etc.", *maa # moo* "for", *ko* "focus", and so on.

The category of possession is marked by particles and possessive pronouns. The nouns themselves cannot be divided into alienable and inalienable since the category specifies the attitude of the possessor to what is possessed: *tooku uupoko* (inalienable) is my head as part of my body while *taaku uupoko* (alienable) may be the head of an enemy killed by myself in a battle. New nouns are created by composition (very productive), derivation (much less productive; the only productive quasi-prefix is *kai-* "actor", e.g. *whakaako* "to teach" - *kaiwhakaako*

"teacher") or by the addition of the nominalizing suffix corresponding roughly to English *-ing* because Maori deverbatives have either derivative force (e.g. *ako* "to learn" - *akonga* "pupil, student") or are simply a syntactic device (e.g., *I tooku haerenga ki Aaakarana* ..., "When I went to Auckland ...", literally "During my going to Auckland ...").

The class of verbs is likewise uninflected, with the exception of the passive (see above). A set of verbal particles is preposed to the head of verbal phrase to specify its temporal, aspectual, and modal meaning. Only two particles may be defined as temporal, i.e., *i* "imperfect" and *e* "future"; most of them are aspectual, i.e., *kua* "perfective", *ka # kaa* "inceptive", *e ... ana* "durative", and the rest are modal, i.e. *kia* "desiderative", *kei* "lest". The postpositive *ai* is anaphoric. Transitive and intransitive verbs are distinguished only by their compatibility or incompatibility with the direct object (marked by *i* or by *ki*, with a restricted set of psychological verbs). However, a transitive action exerted upon an object is expressed as passive plus the patient without the object marker *i*, e.g., *Inumia te wai*! "Drink that water!", literally "Be drunk that water!". Plural, frequentative, reciprocal or durative action is marked with the aid of reduplication, partial or complete (see above).

Basic syntactic features of Maori may be summed up as follows. Maori is an accusative-nominative language notable for a very frequent occurrence of passive constructions. It has a closed class of inherently intransitive words (statives). The neutral word order in an active verbal sentence is "Verbal predicate + Subject + Accusative particle *i* + Direct object". The word order of a passive sentence is very much the same: "Verbal predicate (in the passive form) + Agentive particle *e* + Actor" or "Verbal predicate (in the passive form) + Subject + Agentive particle e + Actor"; the latter alternative, although equivalent to the former one, is more usual. Another elementary rule requires that signifiant follows signifié, with the exception of attributive pronouns that precede their signifié (e.g. *tooku whare* "my house", *ooku whare* "my houses", *teenei whare* "this house", *eetahi whare* "some houses", or they are discontinuous (e.g., *te whare nei* "this house", *ngaa whare nei* "these houses"). In nominal sentences, the nominal predicate takes up the initial position in the sentence and is followed by the subject, e.g. *He whare pai teenei whare* "This house is a good house". However, when the focus is on the subject, the latter is moved to the initial position and preceded by the focus particle *ko*: *Ko te whare pai teenei whare* "The good house is this one". Another peculiarity of Maori are its negative constructions. They consist of one of the negative bases (coded as Neg) that are of verbal nature, placed initially

and followed by the meaningful verb, e.g. *E haere ana ia* "He is going" - *Kaaore ia e haere ana* "He is not going".

I. Column 9 contains information on the number meanings and is based on Williams' dictionary (Williams 1957), however, the data have been modified in those instances when it seemed to be inevitable. Homonyms are listed separately so that the number coded in Column 9 is an index of polysemy.

J. Column 14 codes occasional stylistic characteristics of the lexemes as D (dialect), Mod (modern) or Col (colloquial).

In this respect as well as in its treatment of borrowings, Williams' (1957) dictionary cannot be regarded as an adequate and exhaustive lexicographic work.

References

Biggs, B. (1961), "The Structure of New Zealand Maori". *Anthropological Linguistics* 3, 3, 1-54.

Biggs, B. (1969), *Let's Learn Maori*. Wellington: A.H. & A.W. Reed.

Hohepa, P. (1967), "A Profile-Generative Grammar of Maori". (Indiana University Publications in Anthropology and Linguistics, *Memoir* 20: IJAL 33.2), Baltimore: Waverley.

Krupa, V. (1968), *The Maori Language*. Moscow: Nauka.

Williams, H.W. (1957), *A Dictionary of the Maori Language*. 6th Ed. Wellington: R.E. Owen, Government Printer.

Rieger, Burghard (ed.)
Glottometrika 13
Bochum: Brockmeyer 1992, 241-286

Bewertung von syntaxen für die beschreibung natürlicher sprachen

Wolf Thümmel

1. Das prinzip der einfachheit ist ein prinzip, das in der linguistik spätestens seit Hjelmslev eine rolle spielt, ohne daß es meines wissens in den seither verflossenen jahren dazu gekommen wäre, diesem prinzip eine gewisse exaktheit und praktikabilität zu verleihen. Bei Hjelmslev (1943: 18) heißt es dazu: "in bezug auf dieses prinzip und nur in bezug darauf kann überhaupt ein sinn in die behauptung gelegt werden, eine bestimmte widerspruchsfreie und vollständige lösung sei richtig und eine andere unrichtig. Als richtig wird diejenige lösung angesehen, die in höchstem grade dem einfachheitsprinzip entgegenkommt". Diese sprechweise suggeriert anzunehmen, daß 'einfachheit' ein komparativer oder (wegen der wendung "in höchstem grade") gar ein metrischer begriff sein soll.

Die vorstellung eines metrischen einfachheitsbegriffs findet sich später auch bei Chomsky (1965: 37). Es handelt sich dabei um das "bewertungsmaß" m (evaluation measure m), d.h. um "a function m such that $m(i)$ is an integer associated with the grammar G_i [d.h. der syntax Σ_i] as its value" (Chomsky 1965: 31).

Sowohl bei Hjelmslev als auch bei Chomsky wird die einfachheit von beschreibungssystemen, von grammatiken (also im engeren sinne: von syntaxen) auf daten bezogen: bei Hjelmslev, weil einfachheit einen teil des sog. empirieprinzips ausmacht, bei Chomsky, weil er das problem der definition von 'einfachheit' in dem problem sieht herauszufinden, "how G_i is determined by D_i for each i" (Chomsky 1965: 38), wobei G_i eine "descriptively adequate grammar [...]" ist und D_i für "primary linguistic data" steht.

Freilich gibt es einen unterschied. Bei Hjelmslev steht einfachheit *neben* vollständigkeit (außerdem widerspruchsfreiheit) der beschreibung der daten; der

grad der einfachheit spielt erst dann eine rolle, wenn das prinzip der widerspruchsfreiheit sowie das prinzip der vollständigkeit, d.h. der adäquaten datenbezogenheit, bei der wahl zwischen konkurrierenden beschreibungssystemen kein ergebnis zeitigt. - Bei Chomsky wird durch den begriff 'einfachheit' der bezug zwischen syntaxen und datenmengen wenn auch nicht erst hergestellt - was sollte es schon heißen, daß eine syntax deskriptiv adäquat ist? -, so doch geregelt.

Bei beiden linguisten findet sich die einschränkende vorstellung, daß unter konkurrierenden beschreibungssystemen jeweils eine als die "optimale" ausgewählt wird: Hjelmslev (1943 : 18): "Als richtig wird diejenige [singular!] lösung angesehen, die in höchstem grade dem einfachheitsprinzip entgegenkommt"; Chomsky (1965: 32): "The device would then select one [!] of these potential grammars by the evaluation measure guaranteed by (v). [(v) specification of a function *m* ...]".[1]

Einen wichtigen punkt, der bei der konstruktion eines einfachheitsmaßes zu beachten ist, hat Chomsky (1965: 38) hervorgehoben: "It is also apparent that evaluation measures of the kinds that have been discussed in the literature on generative grammar cannot be used to compare different theories of grammar; comparison of a grammar from one class of proposed grammars with a grammar from another class, *by such a measure*, is utterly without sense". Ich will hier nicht weiter auf die gründe eingehen, die zu diesem diktum führen, noch will ich hier auf das allgemeinwissenschaftstheoretische problem des theorienvergleichs eingehen. Im folgenden soll gewährleistet sein, daß die zu vergleichenden syntaxen ein und demselben syntaxtyp angehören. Wegen der theoreme, mit denen in der theorie des Artikulators (Blanche Noëlle Grunig 1981) die instabilität von dependenzstrukturen[2] aufgezeigt wird, kommen derartige strukturen hier nicht in betracht und folglich auch nicht syntaxen, die diese strukturen festlegen. Es wird vielmehr um syntaxen gehen, mit denen syntaktische relationen über mengen von syntaktischen komplexen und syntaktischen minimaleinheiten erklärt werden. Syntaktische komplexe nennt man gewöhnlich konstituenten.

Chomsky (1965) weist darauf hin, daß das von ihm selbst avisierte, nicht aber konstruierte und angewandte maß der einfachheit (oder: der bewertung), ja überhaupt "comparing two theories of a language - two grammars of this language - in terms of a particular general linguistic theory" eine angelegenheit ist, die für den aufbau einer "explanatory theory of language" entscheidend sei.[3] Ich will hier meine verwunderung darüber nicht weiter erläutern, daß bei der

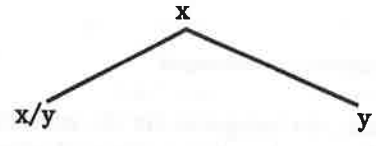
außerordentlichen bedeutung, die man mit Chomsky (1965) einem solchen einfachheitsmaß zuschrieb, das interesse daran nicht so ausgeprägt zu sein schien, daß man sich in den folgejahren ernsthaft damit befaßt hätte.

Selbst wenn man mit einem solchen maß oder mit der idee, auf der die konstruktion eines solchen maßes beruhen könnte, *nicht* unmittelbar den spracherwerb oder die erlernbarkeit in verbindung bringt, muß man sich um ein verfahren kümmern, nach dem zwischen den möglichen syntaxen ein und desselben typs diejenigen ausgewählt werden können, die nach bestimmten festzulegenden prinzipien wenn auch nicht die einfachsten, so doch die optimalen beschreibungen liefern. Hält man ein solches verfahren für überflüssig, so kann man mit der konstruktion von syntaxen zur beschreibung natürlicher sprachen aufhören.

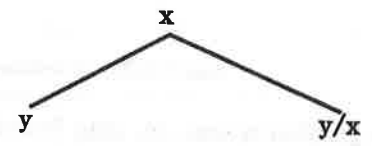
Im folgenden will ich mich um prinzipien kümmern, die bei der konstruktion eines bewertungsmaßes eine rolle spielen könnten.[4] Es wird anhand der ausführungen plausibel werden, daß es nicht so sehr um einfachheit nach genau einem kriterium gehen kann, sondern um eine reihe von prinzipien, die zusammen bei der auswahl "optimaler" syntaxen eine rolle spielen.

Das thema hat in den letzten Jahren durch die elaborierung kategorialsyntaktischer systeme und deren anwendung auf natürlichsprachliche ausdrücke an potentieller aktualität, ja an brisanz gewonnen. Bei der elaborierung handelt es sich um die ausnutzung von funktionenkompositionen. Während klassisch-kanonische kategoriale systeme (Bar-Hillel 1960; in anderer kodierung; Bar-Hillel 1953) (1) oder (2) (oder eine kombination von beidem) als "kürzungsregeln" vorsehen - in baumform als (F1) bzw. als (F2) -, erlaubt funktionenkomposition auch operationen wie (3) und (4) - in baumformen als (F3) bzw. (F4):

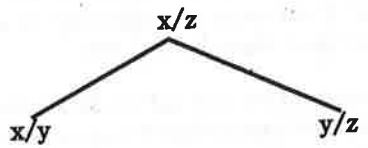
(1) $(x/y).y \Rightarrow x$
(2) $y.(y\backslash x \Rightarrow x$
(3) $(x/y).(y/x) \Rightarrow x/z$
(4) $(z\backslash y).(y\backslash x) \Rightarrow z\backslash x$



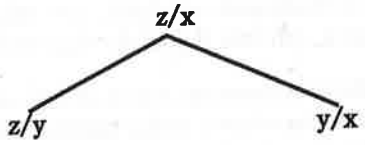
Figur 1: Rechtskürzende kategoriale struktur



Figur 2: Linkskürzende kategoriale struktur



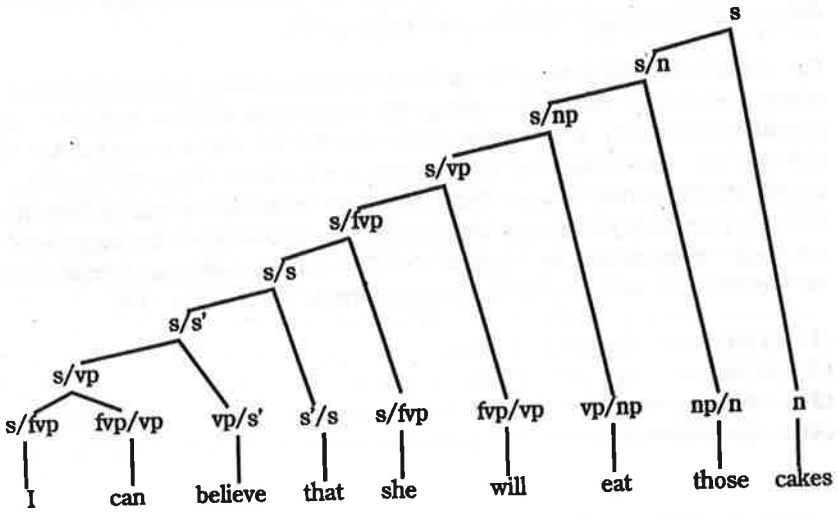
Figur 3: Funktionenkomposition, rechtskürzend



Figur 4: Funktionenkomposition, linkskürzend

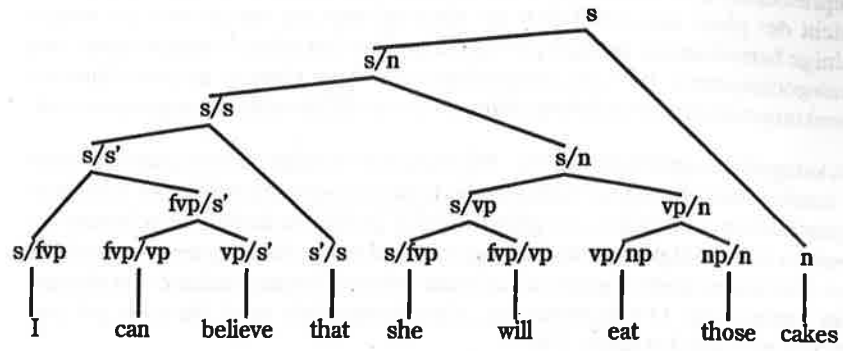
Damit wird es möglich, ein und demselben ausdruck in ein und demselben system eine vielzahl von strukturen zuzuweisen. Steedman (1985: 538) gibt für (5) eine linksverzweigende struktur wie in (F5) an (die vollendung der linksverzweigung stammt von mir (W.Th.)).

(5)I can believe that she will eat those cakes.

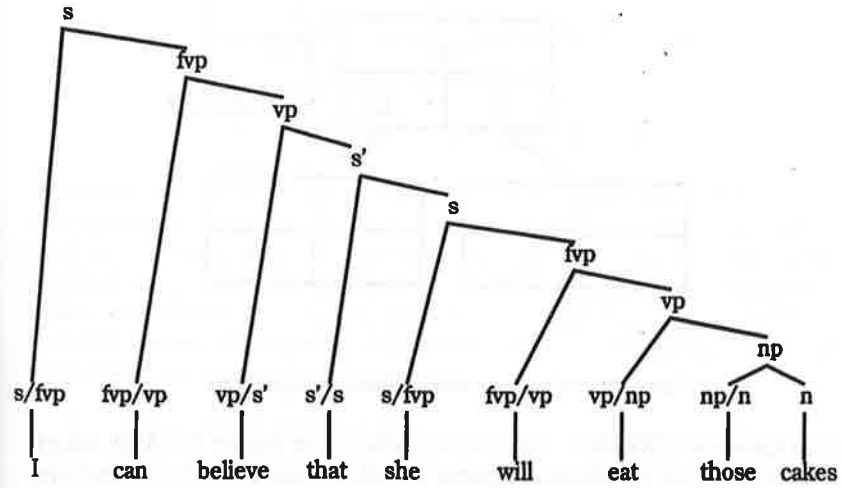


Figur 5: Struktur mit funktionenkomposition, linksverzweigend

Im gleichen system, d.h. unter beibehaltung der kategorien für die atomaren ausdrücke (*I*, *can*, *believe* usw.) können bei anwendung von funktionenkomposition weitere strukturen konstruiert werden, z.b. (F6) oder die rechtsverzweigende struktur (F7), die den vorstellungen der linguistischen tradition entsprechen dürfte.



Figur 6: Alternative struktur mit funktionenkomposition, gemischtverzweigend

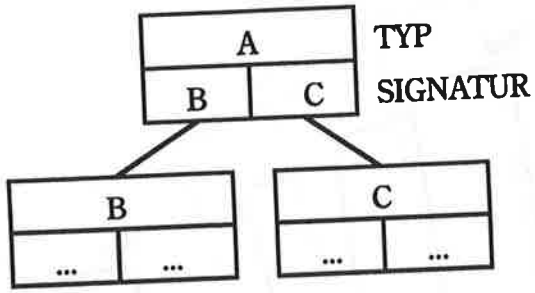


Figur 7: Alternative struktur mit funktionenkomposition, rechtsverzweigend

Steedman (1985: 538) sieht in der vervielfachung von analysen keineswegs ein malheur, vielmehr sagt er: "no reason exists for any autonomous syntactic representation, as distinct from the interpretation itself, to be built." Es ist hier nicht der platz, um ausführlich auf diese behauptung einzugehen. Es mögen einige bemerkungen genügen, um einen eventuellen schaden abzuwenden, den kategorialsysteme für mein unternehmen anrichten können, in denen funktionenkomposition oder andere operationen mit ähnlichen effekten zugelassen sind.

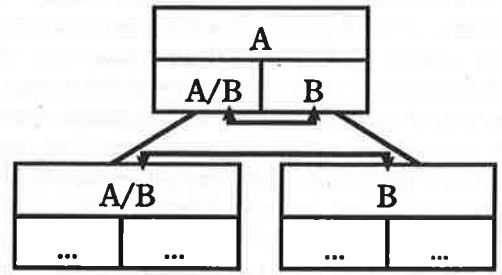
In kategorialen systemen werden - wie in anderen vergleichbaren systemen[4] auch - kombinatorische eigenschaften kodiert, die man den beschriebenen natürlichsprachlichen ausdrücken auf grund welcher analysemethoden auch immer zuweisen will. Lediglich die kodierungsart ist bei kategorialen systemen spezieller art. Der unterschied sei graphisch an hand "lokaler" bäume (an hand von bäumen der kantenlänge 1) veranschaulicht. Eine kontextfreie regel der form (6) entspricht dem lokalen baum (F8):

(6) $A \rightarrow B\ C$

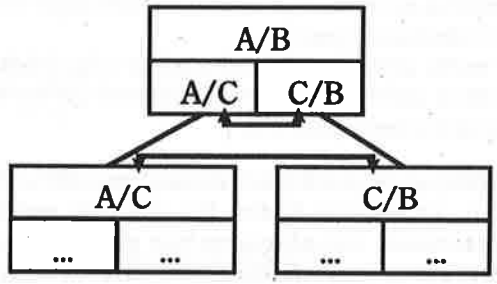


Figur 8: Vertikale kodierung der kombinatorischen eigenschaften

Jedem knoten in (F8) wird - um es in der sprache der theorie des Artikulators auszudrücken - ein typ und eine signatur als etikett zugeordnet. Jeder abschnitt der signatur - in (F8) sind dies: B und C - legt fest, von welchem typ die etiketten der tochterknoten sein müssen. Dabei sind die - im hier gegebenen falle: zwei - abschnitte der signatur voneinander unabhängig. Anders in den kategorialen systemen, wie an den beispielen (F9) und (F10) mit hilfe der pfeile gezeigt sei:



Figur 9: Horizontale kodierung ohne funktionenkomposition



Figur 10: Horizontale kodierung mit funktionenkomposition

Kennt man von den komplexen knotenetiketten in den strukturen (F9) und (F10) den typ und den ersten abschnitt der signatur - A und A/B in (F9); A/B und A/C in (F10) -, so kann man nach dem jeweiligen kürzungsverfahren den zweiten abschnitt der signatur bestimmen. Sind die kombinatorischen eigenschaften in dieser weise kodiert, soll von horizontaler kodierung dieser eigenschaften gesprochen werden. Im folgenden werden nur strukturen und systeme betrachtet, die frei sind von horizontaler kodierung.

2. Die allgemeinen probleme der optimierung von sprachbeschreibungen will ich an einem konkreten fall erörtern. Es soll darum gehen, wie mit hilfe kontextfreier syntaxen niederländische ausdrücke optimal beschrieben werden können, in denen die morphologischen einheiten *en*, *want*, *maar*, *echter* als interne teilausdrücke vorkommen. Daß diese morphologischen einheiten als interne teilausdrücke vorkommen, besagt, daß es jeweils noch einen teilausdruck gibt, der davor steht, und einen, der darauf folgt. Dadurch wird die gän-

gige redeweise verständlich, nach der *en, want, maar, echter* - als "konjunktionen" (voegwoorden) - jeweils paare von ausdrücken miteinander "verknüpfen". Gewöhnlich sagt man, daß es sich dabei um paare von "sätzen" handelt. Um mich hier nicht auf einen bestimmten satzbegriff festzulegen und um nicht unklar formulierte mutmaßungen als vorweggenommenes ergebnis einer in wahrheit nicht durchgeführten untersuchung auszugeben, sage ich, daß das paar jeweils aus einem praecedens und einem subsequens besteht. Ich will es dabei für gerechtfertigt ansehen, unter bestimmten, hier nicht weiter erläuterten bedingungen teilausdrücke, die vor *echter* stehen, nicht zum praecedens, sondern zum subsequens zu rechnen. In (6) gehört danach *Alles is* zum subsequens.

(6) Dit alles wordt aangetoond als een soort documentaire. *Alles is* echter gerenconstrueerd en er wordt gevochten met geestdrift en met een onblusbare vaderlandse geest.
('Das alles wurde als eine art dokumentarfilm ausgegeben, es ist aber alles konstruiert, und es wurde mit enthusiasmus und mit einem unauslöschlichen patriotismus gekämpft.').

Subsequens und praecedens sind folglich abstraktionen, die in der syntaktischen beschreibung ihre entsprechung finden. Der ausdruck von subsequens und praecedens ist systematisch von subsequens bzw. praecedens zu unterscheiden. In diesem artikel gilt diese unterscheidung trotz vereinfachender sprechweise.

Im folgenden verwende ich die formalen mittel kontextfreier syntaxen. Ein hauptgeschäft wird dann sein zu entscheiden, wie den symbolen von konstruierten syntaxen natürlichsprachliche ausdrucksstücke zugeordnet werden sollen; anders gewendet: es ist zu entscheiden, durch welche symbole die natürlichsprachlichen ausdrucksstücke repräsentiert werden sollen. Wenn ich hier von symbolen spreche, orientiere ich mich an den klassischen darstellungen kontextfreier syntaxen: Es handelt sich dabei um das nicht-terminale vokabular. Daß man anstelle solcher symbole komplexe strukturen setzen kann, mit welchen die kombinatorischen eigenschaften kodiert werden können, zeigen nicht nur kategoriale syntaxen, sondern auch systeme wie LFG oder GPSG. Zwar können systeme solcher komplexen strukturen effektiv dazu benützt werden, die kapazität kontextfreier syntaxen erheblich zu beschränken, doch bleiben sie im rahmen dieses artikels unberücksichtigt.

Ich will für das folgende davon ausgehen, daß jedem symbol des inventars - auch jedem terminalen symbol - grundsätzlich eine klasse von natürlichsprachlichen ausdrucksstücken gemäß[5] (a) oder (b) zugeordnet werden kann.

(a) Sie werden als elemente ein und derselben klasse betrachtet und entsprechend einem gemeinsamen symbol des syntaxvokabulars zugeordnet.
(b) Sie werden als elemente verschiedener klassen betrachtet und entsprechend verschiedenen symbolen des syntaxvokabulars zugeordnet.

Somit stellen sich bei der konstruktion geeigneter syntaxen für die beschreibung von ausdrücken, in denen die morphologischen einheiten *en, want, maar, echter* als stücke vorkommen, u.a. folgende fragen:

1. Werden *en, want, maar, echter* wie (a) oder wie (b) behandelt?
2. Werden die praecedentia von *en, want, maar, echter* wie (a) oder wie (b) behandelt?
3. Werden die subsequentia von *en, want, maar, echter* wie (a) oder wie (b) behandelt?
4. Werden die praecedentia der vier morphologischen einheiten zusammen mit den zugehörigen subsequentia jeweils wie (a) oder wie (b) behandelt?

Diese vier fragen sollen im folgenden erörtert werden, indem ich mich an der frage 1 orientiere.

Es sollen folgende annahmen überprüft werden:

en, want, maar, echter

A 1 sind elemente ein und derselben klasse.

{en, wa, ma, ec}

A 2 werden auf zwei disjunktive klassen verteilt.

2.1. {en, wa}, {ma, ec}

2.2. {en, ma}, {wa, ec}

2.3. {en, ec}, {wa, ma}

2.4. {en, wa, ec}, {ma}

2.5. {ma, wa, ec}, {en}

2.6. {ma, en, wa}, {ec}

2.7. {ma, en, ec}, {wa}

A 3 werden auf drei disjunkte klassen verteilt.

3.1. {en, wa}, {ma}, {ec}

3.2. {en, ec}, {wa}, {ma}

3.3. {en, ma}, {wa}, {ec}

3.4. {wa, ec}, {en}, {ma}

3.5. {wa, ma}, {en}, {ec}

3.6. {ma, ec}, {en}, {wa}

A 4 werden auf vier disjunkte klassen verteilt.

{en}, {wa}, {ma}, {ec}

Außer den annahmen A 1 bis A 4 werden auch varianten davon eine rolle spielen, die sich durch folgende beispiele andeuten lassen:

A 1.a {en, wa, ma, ec}; {en, wa}, {en} ... (variante von A 1)

A 2.1.a. {en, wa}, {ma, ec}; {ma, ec, en}, {en} ... (variante von A.2.1)

A 3.3.a. {en, ma}, {wa}, {ec}; {en}, {en}, {ma} ... (variante von A 3.3)

A 4.a. {en}, {wa}, {ma}, {ec}; {ma}, {ma, en}, {ec} ... (variante von A 4).

Diese varianten werfen das problem der syntaktischen mehrdeutigkeit auf.

A 1 ist in verbindung mit der annahme, daß alle praecedentia und alle subsequentia ein und derselben klasse zugeordnet werden, die schwächste und - wenn man will - generellste annahme. Sie legt drei typen von syntaktischen beschreibungen nahe, die u.a. auf den drei verschiedenen syntaxen $\Sigma_{1.1}$, $\Sigma_{1.2}$ und $\Sigma_{1.3}$ beruhen können:

$\Sigma_{1.1}$ = (I = A

V_n = {A, B, K}

V_t = {a_1, ..., a_n; *en*, *wa*, *ma*, *ec*}

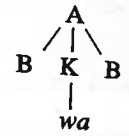
R = {A → B;

A → BKB;

K → *en*, *wa*, *ma*, *ec* ;

B → a_1, ..., a_n})

Beispielstruktur



$\Sigma_{1.2}$ = (I wie bei $\Sigma_{1.1}$

V_n = {A, B, C, K}

V_t = wie bei $\Sigma_{1.1}$

R = {A → B;

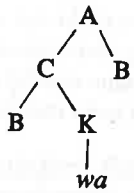
A → CB;

C → BK;

K → *en*, *wa*, *ma*, *ec* ;

B → a_1, ..., a_n})

Beispielstruktur



$\Sigma_{1.3}$ =(I wie bei $\Sigma_{1.1}$

V_n wie $\Sigma_{1.2}$

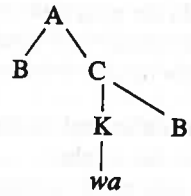
V_t wie $\Sigma_{1.1}$

R = {A → B;

A → BC;

C → KB;

Beispielstruktur



$K \rightarrow en, wa, ma, ec$;

$B \rightarrow a_1, ..., a_n$})

In $\Sigma_{1.1}$, $\Sigma_{1.2}$ und $\Sigma_{1.3}$ werden die praecedentia wie die subsequentia behandelt, und zwar nach dem gesichtspunkt (a). Außerdem liegt A 1 zugrunde.

Mit jeder der drei syntaxen $\Sigma_{1.1}$, $\Sigma_{1.2}$ und $\Sigma_{1.3}$ können ausdrücke des Alge-meen Beschaafd Nederlands (ABN) wie (7) bis (10) beschrieben werden, sofern man $a_1, ..., a_n$ als liste der symbole für alle unanalysierten praecedentia und alle nicht analysierten subsequentia betrachtet.

(7) Ik heb nu gelukkig nergens last van, [= a_1] en [= *en*] wanneer het weer pijn gaat doen neem ik gewoon een paar aspirientjes [= a_2].
('Ich hab nun glücklicherweise nirgends mehr beschwerden, und wenn es mal wieder weh tut, nehme ich gewöhnlich ein paar aspirinchen.')

(8) Er is ongetwijfeld een veelvoud aan portemonnees gerold, [= a_3] want [= *wa*] lang niet iedereen meldt zoiets bij ons [= a_4].
('Da sind ohne zweifel eine menge portemonnaies geklaut worden, denn lange nicht jeder meldet so etwas bei uns.').

(9) Waarschijnljk heeft zij de halfbewusteloze vrouw daarna de gracht in geduwd, [= a_5] maar [= *ma*] deze klemde zich in doodsangst aan haar vast [= a_6].
('Wahrscheinlich hat sie die halbbewutßlose frau danach in die gracht gestoßen, aber diese klammerte sich in ihrer todesangst an ihr fest.').

(10) Afgelopen zondagmorgen deed hij bij de politie aangifte van diefstal van de goedgevulde loterijpot. [= a_7] Hij was daarmee [= $a_{8.1}$] echter [= *ec*] ook de eerste inwoner van Goudswaard die zich zelf ten opzichte van de politie een motief voor de moord op Jannetje Klein verschafte [= $a_{8.2}$].
('Vergangenen Sonntagmorgen erstattete er bei der polizei anzeige wegen diebstahl der gutgefüllten lotterietrommel. Er war damit aber auch der erste einwohner von Goudswaard, der sich selbst gegenüber der polizei ein motiv für den mord an Jannetje Klein verschaffte.').

Als kriterium dafür, welche der drei syntaxen die optimale ist, kann
(a) die zahl *r* der regeln,
(b) die zahl *s* der nicht-terminalen symbole

dienen. Je kleiner *r* und je kleiner *s* ist - so könnte das prinzip lauten -, desto "besser" die syntax. Danach müßte man $\Sigma_{1.1}$ den vorzug geben. Ein ähnliches prinzip werde ich später formulieren (prinzip 5). Dabei werden aber nur solche regeln und nur solche nicht-terminalen symbole zu einem relativen manko einer syntax führen, die nicht systematische konsequenz eines anderen prinzips sind.

Ob ein anderes prinzip das mehr an regeln und symbolen in $\Sigma_{1.2}$ und $\Sigma_{1.3}$ rechtfertigen kann, läßt sich ohne heranziehung neuer daten, anhand derer man die designata von B unter den verschiedenen vorkommensbedingungen untersucht, nicht entscheiden.

Darüber hinaus kann man auch als kriterium die zahl *k* der nicht-terminalen knoten jedes strukturbaumes nehmen, den eine bestimmte syntax einem ausdruckstyp zuzuordnen gestattet. Für ausdrücke (7) bis (10) erhielte man die folgenden werte:

	k (7)	k (8)	k (9)	k (10)
für $\Sigma_{1.1}$	4	4	4	4
für $\Sigma_{1.2}$	5	5	5	5
für $\Sigma_{1.3}$	5	5	5	5

Man kann für jede syntax A einen mittleren knotenindex errechnen, der angibt, wieviele nicht-terminale knoten es in A durchschnittlich je strukturbaum und je ausdruckstyp gibt. Das prinzip könnte lauten:

PRINZIP 1:

Eine syntax A hat gegenüber einer syntax B einen vorteil, wenn ihr mittlerer knotenindex niedriger ist als der von B.

3. Zieht man neue daten heran, so kann man stets die frage stellen, ob diese derselben sprache zuzurechnen sind wie die ausgangsdaten. Ich will für die folgenden darlegungen diese frage nicht für so gravierend halten, daß ich mich durch ihre erörterung davon abhalte, das mir hier gestellte thema zu behandeln. Freilich wird sich zeigen, daß die frage so unbedeutsam nicht ist und letztlich mit dem thema aufs engste zusammenhängt. Man handelt sie gewöhnlich unter dem namenspaar 'homogenität' und 'heterogenität' ab (s. prinzip 9, homogenisierungsprinzip). Ohne heranziehung neuer daten kann eine unwill-

kürliche bevorzugung einer der beiden syntaxen $\Sigma_{1.2}$ und $\Sigma_{1.3}$ überhaupt nicht gerechtfertigt werden, wenn man sich nicht zu mehr oder weniger traditionsorientierten seh- oder sprechweisen verstehen will, die bei alleiniger heranziehung der hier bisher betrachteten datentypen unmotiviert bleiben, etwa: "die konjunktionen *en*, *maar*, *want* (und z.t. *echter*) stehen vor dem zweiten konjunktionsglied" oder "sie stehen nach" dem ersten.

Neue daten, die hier betrachtet werden sollen, sind solche, die belegen, daß imperativformen im praecedens aller vier zu untersuchenden morphologischen einheiten vorkommen können:

(11) Leef matig in alles. [= a_9] En [= *en*] verander uw voeding, vandaag al [= a_{10}].
('Leben Sie in allem maßvoll. Und ändern Sie Ihre ernährung, schon heute.')

(12) Laat u niet in met liefde op het eerste gezicht, [= a_{11}] want [= *wa*] u speelt met vuur! [= a_{12}].
('Lassen Sie sich nicht in eine liebe auf den ersten blick ein, denn Sie spielen mit dem feuer!')

(13) Maak optimistische plannen met voldoende ruimte voor ontplooiing, [= a_{13}] maar [= *ma*] blijf binnen uw eigen mogelijkheden [= a_{14}].
('Machen Sie optimistische pläne mit genügend raum für entspannung, aber bleiben Sie innerhalb ihrer eigenen möglichkeiten.')

(14) Als u tot de mensen behoort die alles meteen kwijt moeten, schrijf dan onmiddelijk uw gedachten op papier. [= a_{15}] Verstuur deze brief [= $a_{16.1}$] echter [= *ec*] nooit [= $a_{16.2}$].
('Wenn Sie zu den menschen gehören, die alles auf einmal loswerden müssen, so bringen Sie Ihre gedanken unmittelbar zu papier. Schicken Sie diesen brief aber niemals ab.')

Anders bei den subsequentia: Während sich in den subsequentia von *en*, *maar* und *echter* imperativformen nachweisen lassen (15, 16, 17) - s. auch (11, 13, 14) -, fehlen belege mit imperativformen im subsequens von *want*.

(15) Verhoog de temperatuur (180° C) [= a_{17}] en [= *en*] laat ze [de kussentjes] 6 tot 7 minuten bakken [= a_{18}].
('Erhöhe die temperatur (180° C) und laß sie [die karamelstückchen] 6 bis 7 minuten backen.')

(16) Het zit er dik in dat u dit jaar uw hart zult verliezen. [= a_{19}] Maar [= *ma*] hoedt u, amoureus en trouwhartig paard! [= a_{20}].
('Es ist sehr damit zu rechnen, daß Sie in diesem jahr Ihr herz verlieren werden. Aber hüten Sie sich, verliebtes und treuherziges pferd!')

(17) Verwarm het frituurvet tot 190° C en leg de kussentjes er weer in: ze beginnen nu te zwellen; [= a_{20}] laat ze [= $a_{21.1}$] echter [= *ec*] nog niet kleuren [= $a_{21.2}$].
('Erwärme das fritierfett auf 190° C und leg die karamelstückchen wieder hinein: sie beginnen nun zu quellen; laß sie aber noch nicht bräunen.')

Berücksichtigt man diese vorkommensbeschränkungen für imperativformen im subsequens von *want*, so liegt es nahe, die hier untersuchten morphologischen einheiten auf mindestens zwei klassen zu verteilen, so daß die annahmen A 2.7, A 3.2, A 3.3, A 3.6 und A 4 in die engere wahl kommen. Daher sind in dieser hinsicht alle syntaxen, mit denen eine dieser annahmen formal ausgedrückt ist, den syntaxen $\Sigma_{1.1}$ und $\Sigma_{1.3}$ überlegen.

Die verteilung auf zwei oder mehr klassen kann zu einer vermehrung der nicht-terminalen symbole und zu einer vermehrung der regeln führen. Diese vermehrung ist aber davon abhängig, daß mit ihr vorkommensbeschränkungen beschrieben werden können, die mit symbol- und regelärmeren syntaxen nicht erfaßt werden.

Die besonderen kombinatorischen eigenschaften, die *want* gegenüber *en*, *maar* und *echter* auszeichnet, kann mindestens in zweierlei weisen in der syntax reflektiert werden:

(a) *want* wird einer anderen klasse zugeordnet als *en*, *maar* und *echter* und folglich in der syntax einem anderen symbol aus V_n.
(b) Das subsequens von *want* wird einer anderen klasse zugeordnet als die subsequentia von *en*, *maar* und *echter* und folglich in der syntax einem anderen symbol als V_n.

Nach diesen gesichtspunkten lassen sich etwa syntaxen wie $\Sigma_{2.1}$, $\Sigma_{2.2}$ und $\Sigma_{2.3}$ konstruieren:

$\boxed{\Sigma_{2.1}}$ =(I = A

V_n = {A, B, D, Keme, Wa}

V_t = {b_1, ..., b_n; d_1, ..., d_m; *en*, *wa*, *ma*, *ec*}

R = {A → B Wa D

A → B

A → D

A → B Keme B

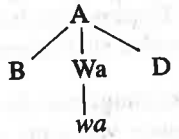
B → b_1, ..., b_n, d_1, ..., d_m

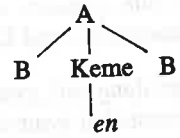
D → d_1, ..., d_m

Wa → *wa*

Keme→ *en*, *ma*, *ec* })

Beispielstrukturen





$\boxed{\Sigma_{2.2}}$ = (I = A

V_n = {A, B, C, D, E, Keme, Wa}

V_t = {b_1, ..., b_n; d_1, ..., d_m; *en*, *wa*, *ma*, *ec*}

R = {A → C D

A → E B

A → B

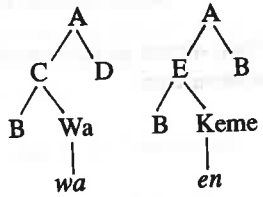
A → D

C → B Wa

E → B Keme

B → b_1, ..., b_n, d_1, ..., d_m

Beispielstrukturen



D → d_1, ..., d_m

Wa → *wa*

Keme→ *en*, *wa*, *ec* })

$\boxed{\Sigma_{2.3}}$ =(I = A

V_n wie $\Sigma_{2.2}$

V_t wie $\Sigma_{2.2}$

R = {A → B C

A → B E

A → B

A → D

C → Wa D

E → Keme B

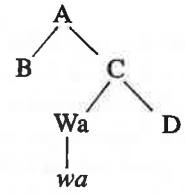
B → b_1, ..., b_n, d_1, ..., d_m

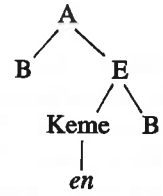
D → d_1, ..., d_m

Wa → *wa*

Keme→ *en*, *wa*, *ec* })

Beispielstrukturen





Mit $\Sigma_{2.1}$, $\Sigma_{2.2}$ und $\Sigma_{2.3}$ werden die praecedentia und die subsequentia voneinander verschieden behandelt: die praecedentia nach dem gesichtspunkt (a), die subsequentia nach dem gesichtspunkt (b).

Die bevorzugung der syntaxen $\Sigma_{2.1}$, $\Sigma_{2.2}$ und $\Sigma_{2.3}$ gegenüber den syntaxen $\Sigma_{1.1}$, $\Sigma_{1.2}$ und $\Sigma_{1.3}$ kann man als anwendungsfall eines allgemeinen prinzips der syntaxbewertung auffassen:

PRINZIP 2:

Syntaxen, in denen beobachtbare vorkommensbeschränkungen von ausdrucksstücken eine formale entsprechung finden, haben solchen gegenüber einen vorteil, in denen dies nicht der fall ist.

Es wäre borniert, wollte ich behaupten, daß sich alle beobachtbaren vorkommensbeschränkungen einzig mit den mitteln kontextfreier syntaxen erfassen lassen. Es wäre aber ebenso borniert, wollte man - aus welchen gründen auch immer - darauf verzichten, vorkommensbeschränkungen mit diesen mitteln zu beschreiben, wenn sie sich mit ihnen ohne schwierigkeiten beschreiben lassen.[6]

Man könnte gegen das prinzip 1 und die damit begründete bevorzugung der syntaxen $\Sigma_{2.1}$ und $\Sigma_{2.3}$ gegenüber den syntaxen $\Sigma_{1.1}$ und $\Sigma_{1.3}$ ins feld führen, daß $\Sigma_{1.1}$ bis $\Sigma_{1.3}$ die allgemeineren syntaxen sind und daß die beobachteten vorkommensbeschränkungen von imperativformen und *want* zwar nicht vernachlässigt, aber mit anderen formalen mitteln erfaßt werden sollen. Gerade dies zu tun habe ich mir jedoch versagt, weil ich nur syntaxen vom gleichen typ miteinander vergleichen will.

4. Mit den bisher konstruierten syntaxen können ausdrücken wie (18) bis (20) zwar strukturbeschreibungen zugeordnet werden, sofern die praecedentia als $a_1, ..., a_n$ und die subsequentia als $a_1, ..., a_n$ bzw. $b_1, ..., b_m$ symbolisiert werden. Da in (18) bis (20) aber je zwei elemente aus K bzw. *wa* mit einem element aus Keme manifestiert sind, bleiben die mit ⌈, ⌉ ausgezeichneten praecedentia von *en*, *maar* und *echter* mit *want* bzw. die mit ⌊, ⌋ ausgezeichneten subsequentia von *want* mit *en*, *maar*, *echter* durch die syntaxen uanalysiert.

(18) ⌈Ontdooi bevroren groenten NIET voor het koken, [= praecedens von *want*] want [= *wa*] ⌊hierdoor worden ze te waterachtig⌉ [= pracedens von *en*] en [= *en*] schorseneren zijn op hun best wanneer ze nog knappend zijn [= subsequens von *en*]⌋ [= subsequens von *want*].
('Taue tiefgefrorenes gemüse nicht vor dem kochen auf, denn hierdurch wird es zu wässrig, und schwarzwurzeln sind am besten wenn sie noch knackig sind.')

(19) ⌈De beste periode is dus van half tot eind april, [= praecedens von *want*] ⌊dan staat alles op zijn mooist,⌉ [= praecedens von *maar*] maar [= *ma*] dan is het op Keukenhof natuurlijk ook het drukst [= subsequens von *maar*]⌋ [= subsequens von *want*].
('Die beste periode ist also von mitte bis ende april, denn dann ist alles am schönsten, aber dann ist in Keukenhof natürlich auch am meisten los.')

(20) ⌈Het wondermiddel zou dus niets nieuws zijn. [= praecedens von *want*] Want [= *wa*] ⌊ook u vindt dat de fiskus recht heeft op uw vrijwillige schenkingen.⌉ [= praecedens von *echter*] ⌈De Fransen zijn het daarmee echter [= *ec*] ⌈slechts indirekt mee eens⌉₂ [= subsequens von *echter*]⌉₁ [= subsequens von *want*].
('Das wundermittel dürfte somit nichts neues sein. Denn auch Sie finden, daß der fiskus ein recht auf Ihre freiwillige spende hat. Die franzosen aber sind damit nur indirekt einverstanden.')

Syntaxen, in denen analysierbare ausdrücke oder ausdrucksstücke nur in ihrer unanalysierten gesamtheit eine formale entsprechung finden, sind gröber als solche, in denen die analysierten ausdrucksstücke eine formale entsprechung finden. Zwei syntaxen können sich also darin unterscheiden, daß die eine feiner, die andere gröber ist.

Dieser gedankengang führt zu:

PRINZIP 3:

Feinere syntaxen haben gegenüber gröberen einen vorteil.

Feinere syntaxen können sich dadurch von gröberen unterscheiden, daß sie mehr terminalsymbole aufweisen als diese. Es kann also der fall auftreten, daß einer syntax A gegenüber einer syntax B der vorzug gegeben wird, weil A mehr terminalsymbole hat, und zwar derart, daß A feiner ist als B.

5. Um ausdrücke wie (18) bis (20) angemessener, d.h. ohne daß das genannte manko auftritt, zu beschreiben, könnte man syntaxen wie $\Sigma_{2.4}$, $\Sigma_{2.5}$, $\Sigma_{2.6}$, $\Sigma_{2.7}$, $\Sigma_{2.8}$, $\Sigma_{2.9}$ u.a. in erwägung ziehen:

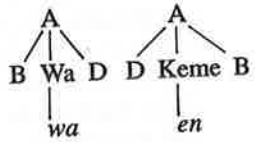
$\Sigma_{2.4}$ =(I = A

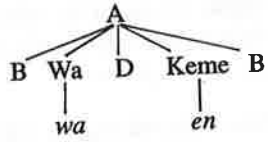
V_n = {A, B, D, Wa, Keme}

V_t = {b_1, ..., b_n; d_1, ..., d_m; *wa*, *en*, *ma*, *ec* }

R = {A → B Wa D
A → D Keme B
A → B Wa D Keme B
A → B
A → D
B → b_1, ..., b_n, d_1, ..., d_m
D → d_1, ..., d_m
Wa → *wa*
Keme→ *en*, *wa*, *ec* })

Beispielstrukturen





In $\Sigma_{2.4}$ findet das praecedens von *want* einerseits und das von *en*, *echter*, *maar* andererseits unterschiedlichen formalen ausdruck. Das praecedens von *en*, *echter*, *maar* wird darüber hinaus nicht einheitlich repräsentiert (bald als D, bald als B Wa D). Entsprechendes gilt für das subsequens von *want* einerseits und das subsequens von *en*, *echter*, *maar* andererseits (bald als D, bald als D Keme B).

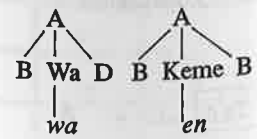
$\Sigma_{2.5}$ =(I = A

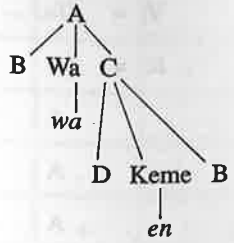
V_n = {A, B, C, D, Wa, Keme}

V_t = {b_1, ..., b_n; d_1, ..., d_m; *wa*, *en*, *ma*, *ec* }

R = {A → B Wa C
A → B
A → D
A → B Wa D
A → B Keme B
C → D Keme B
B → b_1, ..., b_n, d_1, ..., d_m
D → d_1, ..., d_m
Wa → *wa*
Keme→ *en*, *wa*, *ec* })

Beispielstrukturen





$\Sigma_{2.5}$ weist insofern die gleiche art von nachteil gegenüber entsprechenden vergleichssyntaxen auf, als das praecedens von *en*, *echter*, *maar* bald als D, bald als B und das subsequens von *want* bald als C, bald als D, also je verschieden beschrieben werden.

Mit $\Sigma_{2.4}$ werden die praecedentia und die subsequentia, die in gleichen umgebungen stehen, aber teils durch ein nicht- terminales symbol (als kategorie), teils durch eine kette von mehreren nicht-terminalen symbolen beschrieben. Man kann darin einen nachteil von $\Sigma_{2.4}$ gegenüber $\Sigma_{2.5}$ sehen.

Diese uneinheitlichkeit ist ganz offensichtlich ein nachteil von $\Sigma_{2.4}$, wenn man sie mit syntaxen vergleicht, in denen distributionsklassen einheitlich repräsentiert sind. Sinnvoll erscheint somit:

PRINZIP 4:

Syntaxen, die gleiche beobachtbare vorkommensbeschränkungen formal einheitlich beschreiben, haben gegenüber solchen, die diese vorkommensbeschränkungen formal uneinheitlich repräsentieren, einen vorteil.

6. Als weitere alternativsyntaxen sei $\Sigma_{2.6}$ mit den varianten $\Sigma_{2.6.1}$, $\Sigma_{2.6.2}$, $\Sigma_{2.6.3}$ betrachtet:

$\boxed{\Sigma_{2.6}}$ =(I = A

V_n = {A, B, C, D, E, F, Wa, Keme}

V_t = {b_1, ..., b_n; d_1, ..., d_m; *wa*, *en*, *ma*, *ec* }

R = {A → C B

A → B Wa D

A → B

A → D

A → E

C → E Keme

C → B Keme

E → F D

F → B Wa

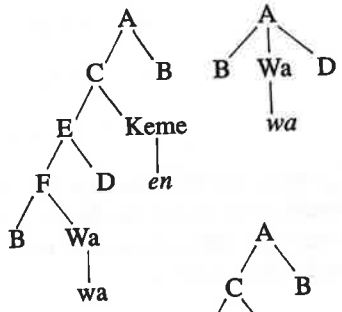
B → b_1, ..., b_n, d_1, ..., d_m

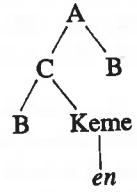
D → d_1, ..., d_m

Wa → *wa*

Keme→ *en*, *ma*, *ec* })

Beispielstrukturen





	$\Sigma_{2.6}$	$\Sigma_{2.6.1}$	$\Sigma_{2.6.2}$	$\Sigma_{2.6.3}$
1. A → CB	+	-	+	-
2. A → B Wa D	+	+	+	+
3. A → B	+	+	+	+
4. A → D	+	+	+	+
5. A → E	+	-	-	-
6. A → B Keme B	-	+	-	+
7. A → E Keme B	-	+	-	-
8. A → F D Keme B	-	-	-	+
9. A → F D	-	-	+	+
10. C → B Keme	+	-	+	-
11. C → E Keme	+	-	-	-
12. C → F D Keme	-	-	+	-
13. E → F D	+	+	-	-
14. F → B Wa	+	+	+	+

Tabelle 1: Die syntaxen $\Sigma_{2.6}$, $\Sigma_{2.6.1}$, $\Sigma_{2.6.2}$ und $\Sigma_{2.6.3}$

Die syntaxen $\Sigma_{2.6}$, $\Sigma_{2.6.1}$, $\Sigma_{2.6.2}$ und $\Sigma_{2.6.3}$ sind untereinander schwach äquivalent. Die zahl der nicht-terminierenden regeln in $\Sigma_{2.6}$ ist gröber (neun) als in den alternativen zu $\Sigma_{2.6}$. Außerdem unterscheiden sich diese syntaxen dadurch, daß

$\Sigma_{2.6.1}$ ohne C,
$\Sigma_{2.6.2}$ ohne E und
$\Sigma_{2.6.3}$ ohne C und ohne E

auskommt. Nimmt man die schwache äquivalenz als beurteilungsgrundlage, so läßt sich sagen, daß die symbole C und E offenbar für die mit den verglichenen syntaxen festgelegte sprache nicht erforderlich sind.

Ersetzt man darüber hinaus $\Sigma_{2.6}$ durch $\Sigma_{2.6.1}$, indem man regel 5 aus $\Sigma_{2.6}$ streicht, so erweist sich nicht nur das symbol E als überflüssig, sondern die gesamte regel 5.

Symbole und regeln, für die dies zutrifft, sollen entbehrlich heißen. Es ist offenkundig, daß entbehrliche symbole und regeln nicht dazu beitragen, eine syntax "besser" zu machen als eine andere. Schließt man nicht von vornherein aus, daß es fälle geben kann, in denen man entbehrliche symbole und regeln in kauf zu nehmen bereit ist, so führt diese überlegung zu einem weiteren prinzip:

> **PRINZIP 5:**
>
> Syntaxen mit weniger entbehrlichen symbolen und regeln haben gegenüber solchen mit mehr entbehrlichen symbolen bzw. regeln einen vorteil.

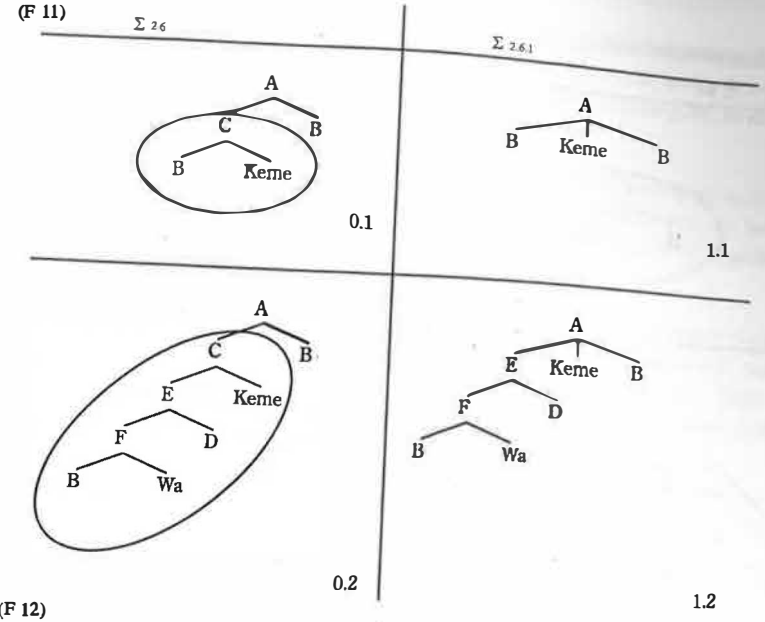
Dieses prinzip sollte man besser in zwei prinzipien zerlegen, weil die zahlen der symbole - anders als es bei $\Sigma_{2.4}$ gegenüber $\Sigma_{2.5}$ der fall ist - sich von denen der regeln unterscheiden können:

> **PRINZIP 5a:**
>
> Syntaxen mit weniger entbehrlichen symbolen haben gegenüber solchen mit mehr entbehrlichen symbolen einen vorteil.

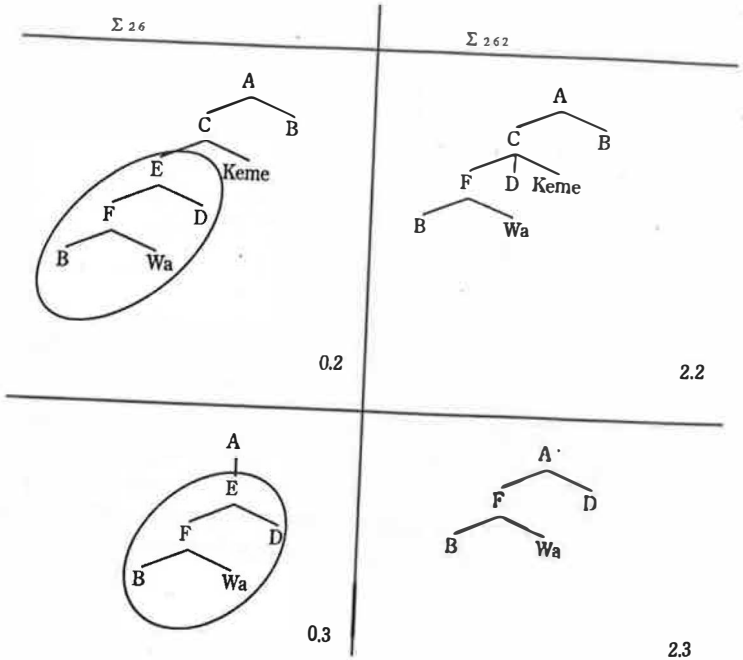
> **PRINZIP 5b:**
>
> Syntaxen mit weniger entbehrlichen regeln haben gegenüber solchen mit mehr entbehrlichen regeln einen vorteil.

7. An hand der syntaxen $\Sigma_{2.6}$, $\Sigma_{2.6.1}$, $\Sigma_{2.6.2}$ und $\Sigma_{2.6.3}$ können weitere betrachtungen angeschlossen werden. Man vergleiche die strukturbäume in (F11) bis (F13).

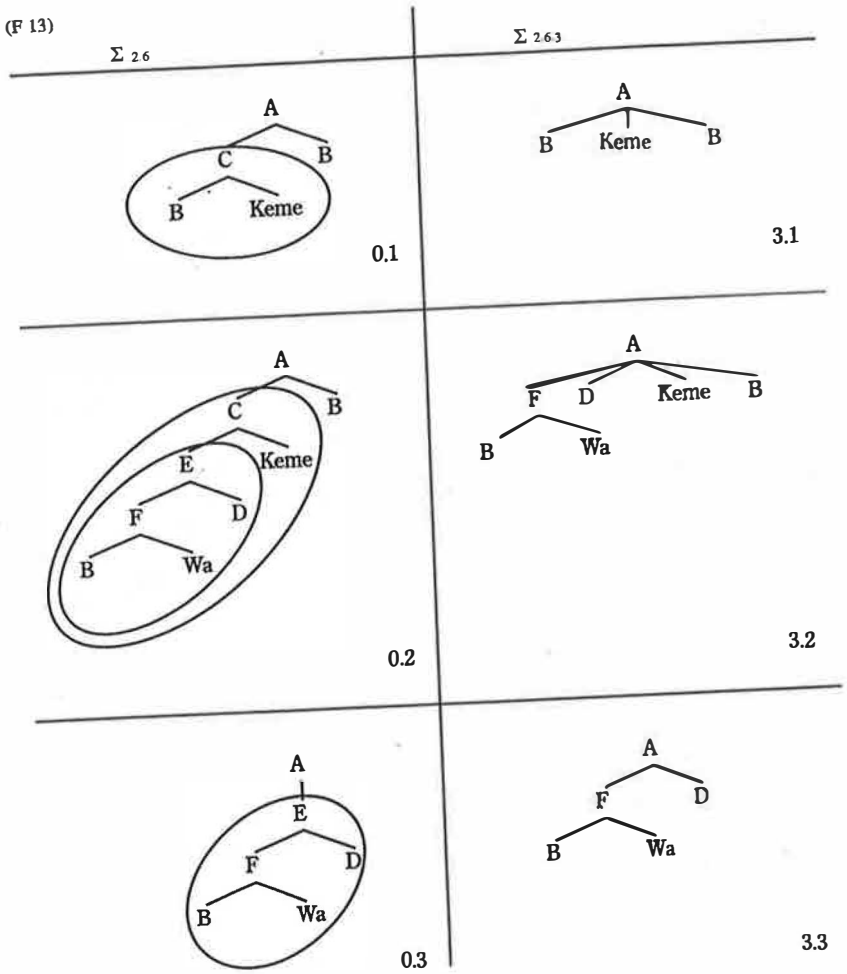
(F 11)



(F 12)



(F 13)



Der zusammenhang, der in (F11) zwischen (0.1) und (0.2) durch das symbol C hergestellt wird, kann mit $\Sigma_{2.6.1}$ nicht ausgedrückt werden, wie (1.1) und (1.2) zeigen. Ebensowenig kann der zusammenhang zwischen (0.2) und (0.3) in (F12) mit $\Sigma_{2.6.2}$ ausgedrückt werden; die entsprechenden strukturen sind (2.2) und (2.3). Entsprechendes gilt in (F13) für (0.1), (0.2) und (0.3) gegenüber (3.1), (3.2) und (3.3).

Wenn es in einer kontextfreien syntax mindestens eine nicht-terminierende regel gibt, in der ein symbol A ∈ Vn auf der linken seite vorkommt, so heiße A '*i*-gradig links-generalisierend', wenn es $i > 1$ nicht-terminierende regeln gibt, in denen A auf der rechten seite vorkommt.

Wenn es in einer kontextfreien syntax mindestens eine nicht-terminierende regel gibt, in der ein symbol A ∈ Vn auf der rechten seite vorkommt, so heiße A '*j*-gradig rechts-generalisierend', wenn es $j > 1$ nicht-terminierende regeln gibt, in denen A auf der linken seite vorkommt.

Es ist nicht abwegig anzunehmen, daß die relative "güte" einer syntax nicht nur mit der zahl links- bzw. rechts-generalisierender symbole wachsen soll, sondern auch in abhängigkeit von der zahl *i* bzw. *j* jedes dieser symbole. - Man kann aufgrund dieser überlegung ein maß *v* der verbundenheit von syn-taxen konstruieren und ein weiteres prinzip aufstellen:

PRINZIP 6:

Syntaxen mit höherem wert für *v* haben gegenüber solchen mit niedrigerem wert für *v* einen vorteil.

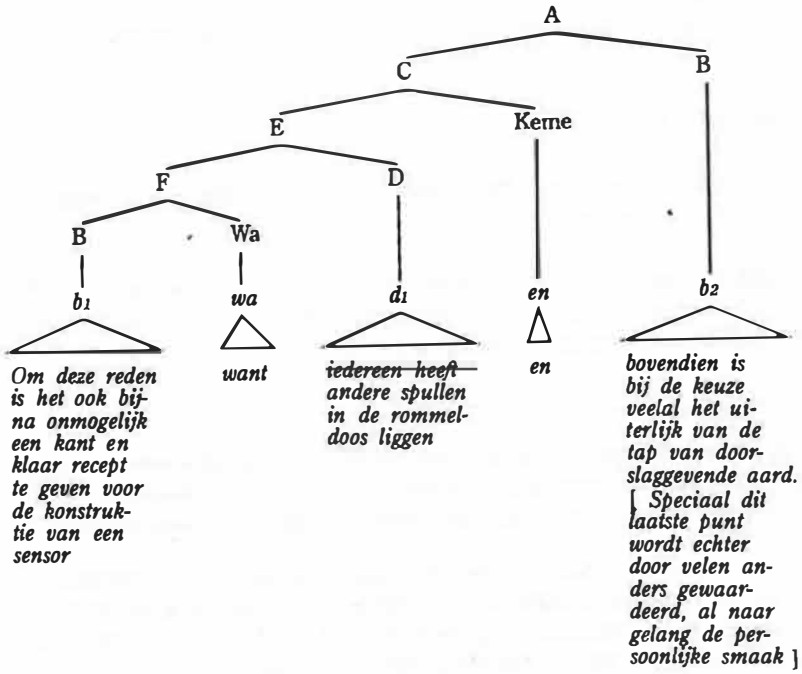
Ein maß *v* werde ich in diesem artikel nicht entwickeln. Wie immer die hier skizzierte idee ausgeführt wird, $\Sigma_{2.6}$ ist aufgrund des prinzips 6 gegenüber allen bisherigen vergleichssyntaxen der vorzug zu geben.

8. Mit $\Sigma_{2.6}$ kann man ausdrücken wie (21) zwar strukturbeschreibungen wie (F14) oder (F15) zuordnen, bei beiden strukturen bleibt allerdings ein ausdrucksstück unanalysiert, das in der je anderen struktur eine analyse erfährt.

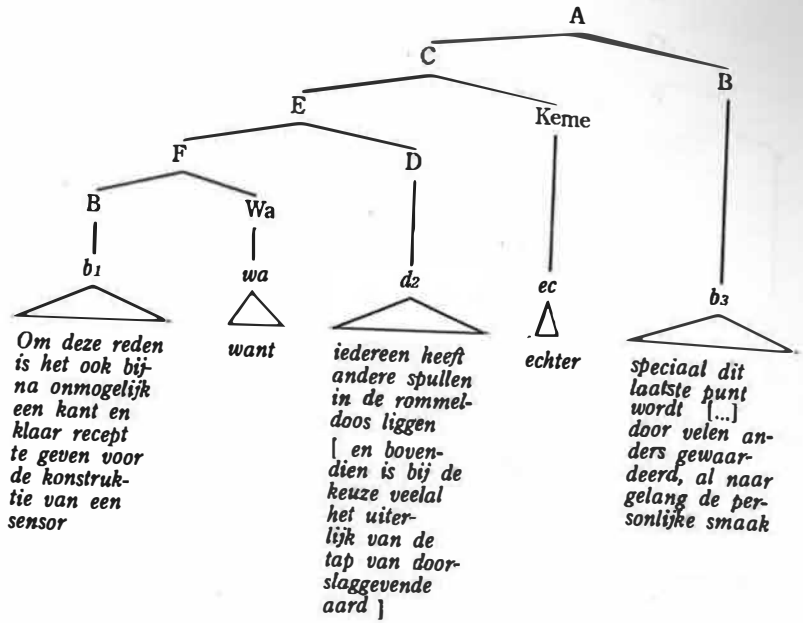
(21) Om deze reden is het ook bijna onmogelijk een kant en klaar recept te geven voor de konstruktie van een sensor, *want* iedereen heeft andere spullen in de rommeldoos liggen *en* bovendien is bij de keuze veelal het uiterlijk van de tap van doorslaggevende aard. Speciaal dit laatse punt wordt *echter* door velen anders gewaardeerd, al naar gelang de persoonlijke smaak.

('Aus diesen gründen ist es auch beinahe unmöglich, ein fix und fertiges rezept für die konstruktion eines sensors zu geben, denn jeder hat andere sachen in seiner schublade, und außerdem spielt bei der wahl meistens die äußere gestalt des stifts eine ausschlaggebende rolle. Besonders dieser letzte punkt wird aber von vielen anders beurteilt, je nach dem persönlichen geschmack.')

(F 14)



(F 15)



Nach dem prinzip 3 wäre gegenüber $\Sigma_{2.6}$ einer syntax der vorzug zu geben, in der die feineren analysen eine formale entsprechung finden. Von den vielen denkbaren syntaxen ist $\Sigma_{2.7}$ ein derartiges exemplar:

$\boxed{\Sigma_{2.7}} = (I = A$

$V_n = \{A, B, C, D, E, F, G, Wa, Keme\}$

$V_t = \{b_1, ..., b_n; d_1, ..., d_m; wa, en, ma, ec\}$

$R = \{A \rightarrow G$

$A \rightarrow G\ F$

G → B

G → B C

G → B F

F → Keme B

C → Wa E

E → D

E → D F

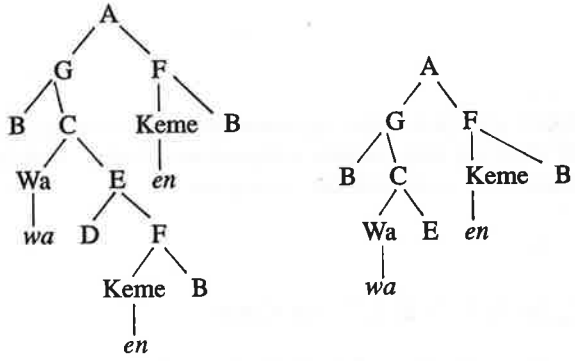
B → $b_1, ..., b_n, d_1, ..., d_m$

D → $d_1, ..., d_m$

Wa → *wa*

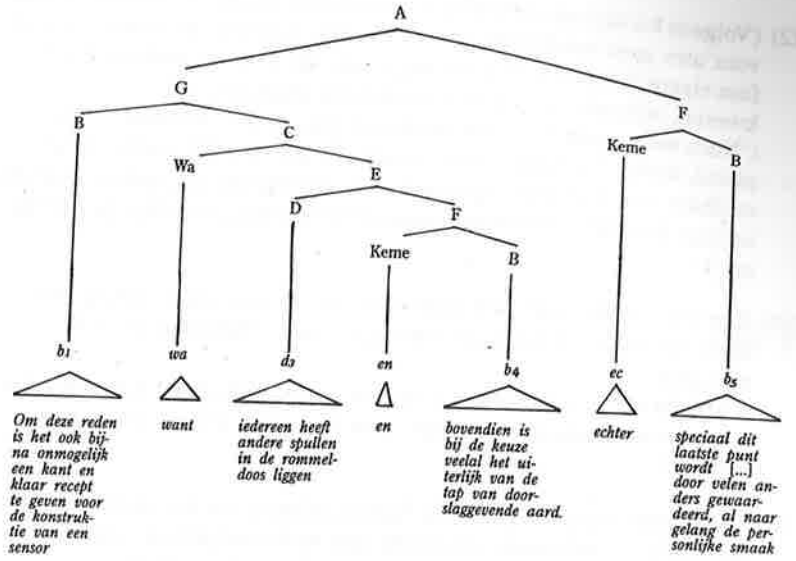
Keme → *en, ma, ec*})

Beispielstrukturen



Mit $\Sigma_{2.7}$ kann dem ausdruck (20) die strukturbeschreibung (F16) gegeben werden:

(F 16)



Mit $\Sigma_{2.7}$ kann ausdrücken wie (21) zwar eine strukturbeschreibung zugeordnet werden, das in runden klammern stehende ausdrucksstück bleibt dabei jedoch unanalysiert.

(21) (Ik schaamde me dood. Maar juist daardoor maakte ik het nog erger), [= b_6] *want* [= *wa*] ik stapte op de juffrouw toe [= d_4] *en* [= *en*] ik fluisterde: 'Juffrouw, ik weet van niks!' [= b_7]. *Maar* [= *ma*] ze begon ontzettend te lachen en met Douwe jr. te stoeien [= b_8].
('Ich schämte mich zu tode. Aber gerade dadurch machte ich es noch schlimmer, denn ich trat an das fräulein heran, und ich flüsterte: 'fräulein, ich weiß von nichts!' Aber sie fing an, fürchterlich zu lachen und mit Douwe jr. zu schäkern.')

Entsprechend bleiben auch die in runden klammern stehenden ausdrucksstücke von (22) bis (24) durch $\Sigma_{2.7}$ unanalysiert.

(22) (Volgens het smeuïg verhaal ging het om 200 nieuwe werkplaatsen waarvoor men geen kandidaten vond *en* de BRT wist later te vertellen, dat de fout eigenlijk bij de werkgever lag,) *want* die bood de werknemers in kwestie, oninteressante, kort lopende kontrakten aan.
('Nach der geschmackvollen erzählung ging es um 200 neue arbeitsplätze, wofür man keine kandidaten fand, und die BRT wußte später zu erzählen, daß der fehler eigentlich beim arbeitgeber lag, denn der bot den in frage kommenden arbeitnehmern uninteressante, kurzfristige verträge an.')

(23) (Op het vlak der resultaten spelen onze deelnemers maar weinig mee. *Maar* het nationaal peil gaat omhoog), *want* de Belgische records tuimelen.
('Auf der ebene der ergebnisse spielen unsere teilnehmer nur wenig mit. Aber das nationale niveau steigt, denn die belgischen rekorde purzeln nur so.')

(24) (Vervolgens gingen zij richting Spanje. [absatz] Ver kwamen de jeugdige avonturiers *echter* niet), *want* even voor de Moerdijkbrug, werden zij ontdekt door een surveillancewagen van de Dordtse politie.
('Danach gingen sie in richtung Spanien. [absatz] Weit kamen die jugendlichen abenteurer aber nicht, denn unmittelbar vor der Moerdijkbrücke wurden sie durch einen streifenwagen der dordtschen polizei entdeckt.')

Darüber hinaus entspricht die beziehung zwischen B und D in $\Sigma_{2.7}$ sowie in den übrigen bisher betrachteten syntaxen nicht der beziehung, wie man sie zwischen den typen von ausdrucksstücken feststellen kann, die diesen symbolen zugeordnet werden sollen. Zwar findet in $\Sigma_{2.7}$ - ebenso in $\Sigma_{2.6}$ - die vorkommensbeschränkung von imperativformen eine formale entsprechung, unberücksichtigt bleibt jedoch, daß die B-ausdrücke $b_1, ..., b_n$ und die D-ausdrücke $d_1, ..., d_m$ der bisherigen syntaxen gemeinsamkeiten aufweisen. So können ganz offensichtlich stücke von B-ausdrücken als stücke von D-ausdrücken auftreten und umgekehrt, u.a. nicht-imperativische finite verbformen. Man könnte ausdrücken wie (22) bis (24) feinere strukturbeschreibungen zuordnen und obendrein die beziehung zwischen B und D in der gewünschten weise festlegen, indem man $\Sigma_{2.8}$ konstruiert:

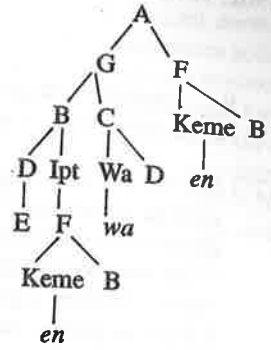
$\boxed{\Sigma_{2.8}}$ =(I = A

V_n = {A, B, C, D, E, F, Ipt, Wa, Keme}

V_t = {$e_1, ..., e_p, ipt_1, ..., ipt_q$, *en*, *ma*, *wa*, *ec* }

R = {A → G
A → G F
F → Keme B
G → B
G → B C
C → Wa D
B → D
B → D Ipt
D → E
D → E F
E → $e_1, ..., e_p$
Ipt → $ipt_1, ..., ipt_q$
Wa → *wa*
Keme→ *en*, *ma*, *ec*})

Beispielstruktur



$ipt_1, ..., ipt_q$ stehen für die charakteristika von imperativformen wie: *ga*, *wees*, *schrijf*, *schrijft* ...

Diese Überlegungen führen zu einem weiteren prinzip:

PRINZIP 7:

Syntaxen, in denen partielle übereinstimmungen von ausdrucksstücken eine formale entsprechung finden, haben gegenüber solchen, in denen sie keine formale entsprechung finden, einen vorteil.

Prinzip 7 entspricht einer in der linguistik gängigen vorstellung. Es entspricht der annahme von sog. endozentrischen konstruktionen (Bloomfield 1933/ 1962: 194).

9. Die bisher betrachteten syntaxen haben gemein, daß sie ausdrücken wie (25) keine feine strukturbeschreibung zuzuordnen gestatten.

(25) Dat was het belangrijkste, *want* in de volgende dagen heeft hij niet veel meer te vrezen, *want* de katalaanse Week wordt minder moeilijk dan ze gisteren en eergisteren was.
('Das war das wichtigste, denn in den folgenden tagen hat er nicht mehr viel zu fürchten, denn die Katalanische Woche wird weniger beschwerlich als sie gestern und vorgestern war.')

Es gibt gute gründe dafür, ausdrücke wie (25), ausdrücke also, in denen zwei stücke unmittelbar aufeinander folgen, die je dem symbol C von $\Sigma_{2.8}$ zugeordnet werden sollen, nicht als mit sicherheit zur niederländischen standardsprache (zum ABN) gehörig zu betrachten. Nicht nur die seltenheit solcher ausdrücke spricht für diese annahme, sondern auch die korrekturbereitschaft, die ihnen gegenüber aktive benützer des ABN haben. Dem beleg (25) steht ein - nicht nur in bezug auf *want* korrigierter - beleg aus einer anderen zeitung vom gleichen tage zur seite.

(25) Dat was het belangrijkste, want in de volgende dagen heeft hij niet veel meer te vrezen. De Katalaanse Week immers minder moelijk.

Sofern diese annahme als berechtigt gelten kann, ist jede syntax, die für die zur diskussion stehenden ausdrucksstückfolgen keine strukturbeschreibungen liefert, in dieser hinsicht solchen vorzuziehen, mit denen ihnen feine strukturbeschreibungen zugeordnet werden können.

Mag meine beurteilung von (25) auch strittig sein, so ist doch immerhin die annahme sinnvoll, daß es ausdrücke gibt, die nicht dem ABN angehören. In

bezug auf solche ausdrücke kann man folgendes prinzip formulieren, wobei die durch eine syntax festgelegten strukturen, denen kein ausdruck der zu beschreibenden sprache entspricht, 'überschuß-strukturen' heißen sollen.

PRINZIP 8:

Syntaxen ohne überschuß-strukturen haben gegenüber solchen mit überschuß-strukturen einen vorteil.

Es ist klar, daß die anwendung des prinzips 7 hinreichende klarheit darüber voraussetzt, welche ausdrücke der zu beschreibenden sprache angehören sollen oder nicht. - Je nach dem empirischen verfahren, mit dem ausdrücke auf syntaxen bezogen werden, und je nach der art der datengewinnung und -bearbeitung kann es sich als nützlich erweisen, zwischen zwei sorten von überschuß-strukturen zu unterscheiden: (a) solche, die (noch) nicht belegt werden konnten, von denen aber aus guten gründen belegbarkeit angenommen werden kann, (b) solche, die nicht belegbar sind, weil von ihnen aus guten gründen eine belegbarkeit gar nicht erwartet werden kann.

10. Wie das nebeneinander von beispielen wie (25) und (26) veranschaulicht, ist es denkbar, daß man - willkürlich oder nach bestimmten kriterien - ausdrücke als verschiedenen sprachen angehörig betrachten kann, auch wenn sie partielle - etwa phonologisch beschreibbare - übereinstimmungen aufweisen. Dabei kann es sich darum handeln, daß man zwei (oder mehrere) nicht-pathologische sprachen oder eine (oder mehrere) nicht-pathologische von einer oder von mehreren pathologischen sprachen unterscheiden will. Um der möglichen willkür und der möglichen unzuverlässigkeit der verwendeten kriterien zu steuern, ist es sinnvoll, ein kompensierendes prinzip aufzustellen, das bewirken soll, daß "unnötige" unterscheidungen von sprachen ("varietäten", "dialekten", "registern") weitgehend vermieden werden.

PRINZIP 9:

Eine syntax A hat gegenüber einer syntax B einen vorteil, wenn ein bestimmter ausdruckstyp durch A eine strukturbeschreibung erhält, nicht aber durch B.

Man könnte - ziemlich willkürlich - ausdrücke mit imperativformen einer anderen sprache zurechnen als ausdrücke ohne imperativformen und für beide ausdruckstypen je eine syntax konstruieren: Σ_{+imp} und Σ_{-imp}. Beide syntaxen sind nach gängigem urteil unbefriedigend. Zwar läßt sich nach prinzip 9 nicht entscheiden, welcher der vorzug zu geben ist, weil beide gegenüber der jeweils anderen ein manko aufweisen (mit Σ_{+imp} werden keine indikativischen ausdrücke, mit Σ_{-imp} keine imperativischen ausdrücke beschrieben), vergleicht man beide aber mit $\Sigma_{2.8}$, so muß $\Sigma_{2.8}$ nach prinzip 9 gegenüber den beiden anderen syntaxen der vorzug gegeben werden. $\Sigma_{2.8}$ vereinigt die relativen vorzüge und kompensiert die relativen nachteile der beiden ausgangssyntaxen. Man kann das prinzip 9 'homogenisierungsprinzip' nennen.

11. Die bisher erörterten acht prinzipien sind informell gegeben. Sie dürften freilich in ihrer informellen fassung bereits andeuten, wo die probleme liegen:

Jedes der prinzipien für sich genommen (abgesehen von prinzip 5) könnte auf die eine oder andere weise der festlegung eines "einfachheitsmaßes" dienen. Sollen aber alle acht und mit sicherheit noch weitere prinzipien bei der auswahl der optimalen syntax (oder der optimalen syntaxen) simultan berücksichtigung finden, so ist zu bedenken, daß nicht alle prinzipien notwendigerweise in die gleiche richtung wirken. Die prinzipien sind nämlich offensichtlich z.t. voneinander abhängig. Gerade diese abhängigkeit ist es, die in dem zu konstruierenden bewertungsmaß ihren ausdruck finden muß.

Der sinn, der in der konstruktion eines solchen maßes liegen kann, ist von zweierlei art:

(i) Für den syntaktischen typologischen sprachenvergleich ist es erforderlich, die für jede sprache gesondert zu schreibende syntax (realistischer: gesondert zu schreibende syntaxen) zu standardisieren. Der vergleich selbst beruht dann auf formal standardisierten syntaxen und nicht auf ontologisch schwerbefrachteten trivialsyntaktischen mustern vom typ

SOV SVO VSO

OSV OVS VOS

o.ä., wie sie mitte des letzten jahrhunderts erfunden, von Wundt systematisiert und von Greenberg in der zweiten hälfte des 20. jahrhunderts neu in umlauf gebracht wurden. Erst auf der grundlage standardisierter syntaxen ist ein syntaktischer vergleich verschiedener nicht-pathologischer, aber auch pathologischer und nicht-pathologischer sprachen möglich, etwa auf grund eines auf die standardisierten syntaxen angelegten tiefenmaßes, wie es in der literatur bereits mehrfach erwogen wurde (Yngve 1961: 130ff.; Papp 1966: 77; Altmann & Lehfeldt 1973: 115ff.). Auch syntaktisch erfaßbare entwicklungen von sprachen lassen sich dann einheitlich und vergleichbar darstellen.

(ii) Man kann - wenn man es für angebracht hält - ein solches bewertungsmaß dazu benützen, die mit seiner hilfe ausgewählte syntax - wenn es sich denn um eine handelt - als eine hypothese über "the linguistic intuition of the native speaker", oder gar als " 'acquisition model' for language" (Chomsky 1965: 24ff.) zu betrachten. Allerdings kann mit recht die annahme bezweifelt werden, nach der dadurch "the class of potential grammars [might be] sufficiently limited so that no more than a single [!] permitted grammar will be compatible with the available data at the moment of successful language acquisition" (Chomsky 1965: 36). Es läßt sich zeigen, daß auch die ergänzung solcher konkurrierender grammatiken durch ein bewertungsmaß für die behandlung des spracherwerbs sowie für die rechtfertigung der selektion einer spezifischen grammatik nicht notwendigerweise zu dem erwünschten ziel führt, daß man vielmehr mit einem solchen maß - vernünftig konstruiert - auf bereiche treffen kann, in denen eine sprache syntaktisch "unstabil" ist, d.h. in denen sie durch verschiedene syntaxen mit gleichem wert des bewertungsmaßes beschrieben werden kann (Thümmel 1985). Weder kann in solchen fällen die "innate human *faculté de langage*" oder die erlernbarkeit der sprache erklärt werden, noch lassen sich beide - einzeln oder gemeinsam - als erklärungsrettendes prinzip heranziehen.

12. Die neun prinzipien sind voneinander nicht unabhängig. Prinzip 1 bewertet gewissermaßen den strukturellen "aufwand" für jedes syntaxpaar. Hiervon sind - wenn man sich auf eine darstellung in form von baumgraphen bezieht - in besonderer weise syntaxen betroffen, die folgen unär verzweigender knoten zulassen.[7]

Prinzip 2 läuft darauf hinaus, daß syntaxen mit mehr überschußstrukturen einen nachteil gegenüber solchen mit weniger überschußstrukturen haben. Prinzip 2 entspricht im effekt dem prinzip 8 und kann daher als selbständiges prinzip berücksichtigt bleiben.

Prinzip 3 kommt hauptsächlich in zwei fällen zum zuge. Eine feinere syntax Σ_f wird man - bezogen auf eine gegebene gröbere syntax Σ_g dann konstruieren wollen, wenn

(a) mit Σ_f beobachtbare vorkommensbeschränkungen von ausdrucksstücken eine formale entsprechung finden, mit Σ_g jedoch nicht;
(b) mit Σ_f partielle übereinstimmungen von ausdrucksstücken eine formale entsprechung finden, mit Σ_g jedoch nicht.

Der fall (a) ist bereits durch prinzip 2, der fall (b) durch prinzip 7 abgedeckt.

Ob es fälle gibt, in denen das prinzip 3 aus anderen gründen als den unter (a) und (b) genannten wirksam ist oder wirksam sein sollte, überblicke ich noch nicht. Da das prinzip 3 bis auf derartige fälle, die ich im folgenden vernachlässige, von den prinzipien 2 bzw. 7 abgedeckt ist, kann ich prinzip 3 bei den folgenden überlegungen außerachtlassen.

Prinzip 4 ist eine spezialisierung des prinzips 2. Es wird jedoch nicht wie dieses durch prinzip 8 abgedeckt. Vielmehr steht es in engem zusammenhang mit prinzip 5 und vor allem mit prinzip 6. Einheitliche beschreibung von vorkommensbeschränkungen wird dadurch erzielt, daß man eine syntax mit größerer verbundenheit (höherem wert für v) konstruiert. Prinzip 4 erübrigt sich demnach als eigenständiges prinzip.

Prinzip 5 lasse ich bei den folgenden überlegungen außer betracht, weil es nur indirekt datenbezogen ist und sich als so plausibel aufdrängt, daß es in der regel gar nicht dazu kommen dürfte, syntaxen in der praxis des syntaxschreibens nach diesem prinzip vergleichen zu müssen.

Während durch das prinzip 9 eine "homogenisierung" derart zuwege gebracht wird, daß man die beschreibungen unterschiedlicher ausdrücke oder aus-druckstypen in einer einzigen syntax zu vereinigen trachtet, sollen durch das prinzip 6 triviale fälle solcher homogenisierung ausgeschlossen werden. Prinzip 6 stellt den zusammenhang zwischen den homogenisierten syntaktischen beschreibungen her. Prinzip 7 ist ganz offensichtlich durch prinzip 6 abgedeckt: Die formale entsprechung von partiellen übereinstimmungen von ausdrucksstücken wird erzielt durch ein und dasselbe element aus dem nicht-terminalen vokabular. Dies läuft auf eine verstärkung der links-generalisiertheit dieses symbols und somit auf eine erhöhung des wertes von v hinaus. Außerdem wird dadurch auch die zahl der entbehrlichen symbole sowie die zahl der entbehrlichen regeln (prinzip 5) verringert.

Wie kann man nun ein bewertungsmaß oder ein maß der relativen "güte" konstruieren? Ich will hier andeuten, welchen weg ich für sinnvoll halte, und mich dabei im wesentlichen an den prinzipien 8 und 9 orientieren. Unberücksichtigt bleiben hier die probleme, die die annahme syntaktischer mehrdeutigkeit mit sich bringt.

Es soll vorausgesetzt werden, daß es für jedes paar zu vergleichender syntaxen mindestens einen ausdruckstyp gibt, dem mit jeder der beiden syntaxen mindestens eine strukturbeschreibung zugeordnet werden kann.

Da man nicht weiß, wie eine "beste" syntax aller konkurrierenden syntaxen für die beschreibung einer bestimmten sprache aussieht, ist es sinnvoll, ein maß für den paarweisen vergleich zu konstruieren, wobei darauf zu achten ist, daß für 'besser als' transitivität gewährleistet ist.

Außerdem muß gewährleistet sein, daß es endliche mengen von strukturen sind, die verglichen werden. Es muß daher ein verfahren geben, das auch bei unendlichen formalen sprachen eine endliche teilmenge von strukturen einheitlich auszeichnet. Ich will ein solches verfahren hier voraussetzen und nenne nach bestimmten kriterien gebildete mengen so ausgezeichneter strukturen 'strukturtypen'. Jedem strukturtyp soll ein τ zukommen, z.b. 1 oder 2 oder 0.5. Es kann sinnvoll sein, verschiedene arten von strukturtypen zu unterscheiden und verschieden zu bewerten.[8]

Ausgehend von diesen werten sollen folgende weiteren werte berücksichtigt werden:

Für zwei syntaxen Σ_i und Σ_j:

π(λήρής) summe der τ-werte aller strukturtypen, für die es belege gibt,

χ(ενός) summe der τ-werte aller strukturtypen, für die es keine belege gibt,

$b(onus)_{ij}$ zahl der strukturtypen, die für Σ_i belegt sind, mit Σ_j aber gar nicht oder nur grob beschrieben werden,

$m(alus)_{i,j}$ zahl der strukturtypen, die für Σ_j belegt sind, mit Σ_i aber gar nicht oder nur grob beschrieben werden.

Wenn man zwei syntaxen Σ_i und Σ_j hinsichtlich ihrer relativen güte miteinander vergleicht, so will man jeder von ihnen einen wert g zuweisen. Es ist sinnvoll anzunehmen, daß Σ_i und Σ_j dann von gleicher güte sind, wenn

(i) $\quad m(\Sigma_i, \Sigma_j) = 0 = m(\Sigma_j, \Sigma_i)$

und folglich:

(i') $\quad b(\Sigma_i, \Sigma_j) = 0 = b(\Sigma_j, \Sigma_i)$

und

(ii) $\quad \varkappa(\Sigma_i) = \varkappa(\Sigma_j)$

Außerdem soll bei $\varkappa(\Sigma_i) = 0$ und $m(\Sigma_i, \Sigma_j) = 0$ der wert für $g(\Sigma_i, \Sigma_j)$ gegen 1 gehen, wenn $\pi(\Sigma_j)$ gegen 0 geht.

Weiterhin ist es sinnvoll anzunehmen, daß

(iii) $\quad \pi(\Sigma_i) > 0$ und $\pi(\Sigma_j) > 0$

(iv) $\quad 0 < g(\Sigma_i, \Sigma_j) \leq 1$

Während es sinnvoll ist, die τ-werte verschiedener strukturtypen verschieden zu gewichten, scheint es nicht sinnvoll, bei den b- und m-werten eine entsprechende gewichtung dadurch vorzunehmen, daß diese werte mit bezug auf die τ-werte in den jeweiligen vergleichssyntaxen berechnet werden, wie dies Clément & Thümmel (1979: 114ff.) vorgeschlagen haben (so auch noch Thümmel 1985). Nimmt man als basis für die berechnung der b- und m-werte - wie ich hier vorschlage - die zahl der im vergleich zur jeweiligen bezugssyntax nicht beschriebenen ausdruckstypen, muß man die gewünschte transitivität der gütebeziehung zwischen syntaxen nicht auf umständliche und wenig motivierte weise herstellen.

Nach den bisherigen erörterungen läßt sich für den paarweisen syntax-vergleich ein maß der relativen güte konstruieren, das etwa die form

$$g(\Sigma_i, \Sigma_j) = \frac{\alpha\pi_i}{\alpha\pi_i + \alpha\pi_j + \beta\varkappa_i + \gamma \cdot m(\Sigma_i, \Sigma_j)}$$

hat. α, β und γ sind variablen für evtl. gewünschte gewichtungen.

Literaturnachweis

[Ades & Steedman 1982] Ades, Anthony E. & Steedman, Mark]: *On the order of words.* - In: Linguistics and Philosophy 4 (1982) 517-558.

[Altmann & Lehfeldt 1983] Altmann, Gabriel & Lehfeldt, Werner: *Allgemeine Sprachtypologie. Prinzipien und Meßverfahren.* - München: Fink (1973).

[Bańczerowski 1980] Bańczerowski, Jan: *Systems of semantics and syntax. A determinational theory of language.* - Warszawa & Poznan: Panstwowe Wydawnictwo Naukowe 1980.

[Bar-Hillel 1953] Bar-Hillel, Yehoshua: *A quasi-arithmetical notation for syntactic description.* - In: Language 29 (1953) 47-58.

[Bar-Hillel 1960] -: *On categorial and phrase structure grammars.* - In:Bar-Hillel, Yehoshua: Language and Information. Selected essays on their theory and application. - Reading [etc.]: Addison-Wesley; Jerusalem: The Jerusalem Academic Press (1964). S. 99-115.

[Billroth 1832] Billroth, Johann Gustav Friedrich: *Lateinische Syntax für die oberen Klassen gelehrter Schulen.* - Leipzig: Weismann 1832.

[Bloomfield 1933] Bloomfield, Leonard: *Language.* (Revised and first published in Great Britain 1935. Reprinted 1962). - London: Allan & Unwin (1962).

[Chomsky 1965] Chomsky, Noam: *Aspects of the theory of syntax.* - Cambridge, Massachusetts: The M.I.T. Press (1965).

[Chomsky 1981] -:*Lectures on Government and Binding. [The Pisa lectures].* - Dordrecht & Cinnaminson: Foris 1981.

[Chomsky 1986] -:*Barriers.* - Cambridge, Massachusetts & London: The M.I.T. Press (1986).

[Clark 1866] Clark, S[tephan] W.: *A practical grammar: in which words, phrases, and sentences are classified according to their offices; and their various relations to one another, illustrated by a complete system of diagrams.* - New York: Barnes 1866.

[Clément & Thümmel 1979] Clément, Danièle & Thümmel, Wolf: *Principes sous-jacents à la description syntaxique d'une langue naturelle.* - In: DRLAV. Documentation et Recherche en linguistique allemande contemporaine - Vincennes. Papier N° 19: Questions de méthode à propos de la construction d'une syntaxe: sur l'exemple de l'allemand standard. - Paris: Université de Paris VIII, Centre de Recherche 1979. S. 19-137.

[Cresswell 1973] Cresswell, M[ax].: *Logics and languages.* - London: Methuen (1973).

[Grunig 1981] Grunig, Blanche Noëlle: *Structure sous-jacente: essai sur les fondements théoriques.* - Lille: Université de Lille III 1981.

[Hjelmslev 1943] Hjelmslev, Louis: *Omkring sprogteoriens grundlæggelse.* - In: Festskrift udgivet af Københavns Universitet i Anledning af Universitets Aarsfest November 1943. - København 1943: Luno. S. 3-113.

[Papp 1966] Papp, Ferenc: *On the depth of Hungarian sentences.* - In: Linguistic 25 (1966) 58-77.

[Steedman 1985] Steedman, Mark: *Dependency and coordination in the grammar of Dutch and English.* - In: Language 61 (1985) 523-568.

[Thümmel 1985] -: *Erfassung instabiler konstruktionen mit kontextfreien syntaxen.* - In: Kontextfreie Syntaxen und Verwandtes. Hrsg. von Ursula Klenk. - Tübingen: Niemeyer 1985.

[Yngve 1961] Yngve, Victor H.: *The depth hypothesis.* - In: Proceedings of Symposia in Applied Mathematics. Vol. XII: Structure of Language and its Mathematical Aspects. - Providence: American Mathematical Society 1961. S. 130-138.

Anm. 1: Es kommt mir hier überhaupt nicht auf eine exegese von historischen Chomsky-texten an; am rande sei nur vermerkt, daß einige stellen bei Chomsky (1965) auch die möglichkeit zulassen, mehrere syntaxen als in gleichem maße erklärungsadäquat anzusehen (etwa Chomsky 1965: 203, anm. 22).

Anm. 2: Allgemeiner: von aranealstrukturen, zu denen z.b. auch die sog. determinationsstrukturen bei Banczerowski (1980) gehören. Unter 'aranealstrukturen' sollen alle solchen strukturen verstanden werden, in denen eine zweistellige relation über eine menge atomarer ausdrücke - über einer menge von "minimaleinheiten" in der theorie des Artikulators - (üblicherweise als eine abhängigkeits- oder als eine determinationsrelation gedeutet) erklärt ist. Solche strukturen haben in der linguistik eine lange tradition und finden sich graphisch dargestellt in Europa z.b. bei Billroth (1832), in Amerika z.b. bei Clark (1866).

Anm. 3: Chomskys formulierung ist in zweierlei hinsicht befremdlich: (a) Eine grammatik für eine natürliche sprache als theorie zu bezeichnen, ist nur bei einem weitestgehend liberalen verständnis von 'theorie' erträglich. In einer entwicklungsphase der linguistik, in der die unsitte geduldet wird, schon irgendeine singuläre hypothese oder irgendeine spezielle form der kodierung theorie zu nennen, mag es vielleicht gerade noch statthaft sein, eine syntax wegen ihres deduktiven und wegen des mit ihr angestrebten prognostischen charakters als theoruncula zu betrachten. - (b) Eine theorie wird im allgemeinen stets als erklärungssystem konzipiert, im standardfall als ein axiomatisiertes. Daher ist es gänzlich unbegreiflich, wie erst die annahme eines bewertungsmaßes eine theorie "erklärend" soll machen können. Erklärend ist z.b. eine der wenigen mir bekannten linguistischen theorien, nämlich die bereits erwähnte theorie des Artikulators, und zwar eben deswegen, weil es sich bei ihr um eine theorie im strikten sinne handelt. Und erklärend ist sie, ohne daß in ihr ein bewertungsmaß festgelegt wäre oder gar deduziert werden könnte. Vielmehr muß es so sein, daß die einführung eines bewertungsmaßes kompatibel sein muß mit der theorie, in die es - sei es als eine ihrer komponenten (als axiom oder als theorem) oder als ein auf sie bezogenes methodologisches prinzip - integriert werden soll. - Wird ein bewertungsmaß in der einen oder anderen weise im rahmen der theorie des Artikulators konzipiert, so muß es z.b. kom-

patibel sein mit dem theorem der erhaltung der minimaleinheiten (Grunig 1981: 183ff.; beweis: 220ff.). Da alle bisherigen heilungsversuche - soweit ich sie kenne - gezeigt haben, daß viele syntaktische strukturen, die auf dem hintergrund der lehre von Rektion und Bindung (Chomsky 1980; 1986) vorgeschlagen werden, dieses theorem verletzen, ist es abwegig, ein bewertungsmaß zu konstruieren, das systematisch bezug nimmt auf eigenschaften derartiger strukturen. Im übrigen ist es nutzlos, das bewertungsmaß so zu gestalten, daß strukturen oder ganze syntaxen, die in vergleichbarer weise gegen theoreme der bezugstheorie verstoßen, mit ihm überhaupt noch bewertet werden müssen. Dies gilt jedenfalls so lange, als nicht eines oder mehrere jener axiome über bord geworfen werden, die zum beweis des theorems der erhaltung der minimaleinheiten führen.

Anm. 4: Für förderliche kritische anmerkungen und hinweise auf fehler in einer früheren version dieses artikels danke ich Ursula Klenk und Burghard Rieger.

Anm. 5: Für die in diesem aufsatz erörterten beispiele natürlichsprachlicher ausdrücke und ihnen zugeordneter strukturen sei diese vereinfachung erlaubt. In allgemeinerer hinsicht ist es sinnvoll anzunehmen, daß es sich (a) nicht nur um symbole handelt, sondern - sofern es nichtterminale sind - auch um die von ihnen dominierten teilstrukturen (oder gar um spezielle teile dieser teilstrukturen - "substrukturen" in der theorie des Artikulators (Grunig 1981: 137ff.)), und (b) nicht nur um klassen natürlichsprachlicher ausdrucksstücke, sondern auch um eigenschaften solcher ausdrücke (z.B. umlaut oder intonation).

Anm. 6: Die in diesem artikel herangezogenen daten weisen kombinatorische eigenschaften auf, die mit kontextfreien syntaxen erfaßt werden können, unter der voraussetzung allerdings, daß zwischen der syntaktischen struktur einerseits und der manifestation der struktur andererseits unterschieden wird. Diese unterscheidung, die ihr pendant etwa in der unterscheidung zwischen 'ausdrucksform' und 'ausdruckssubstanz' bei Hjelmslev (1943), in der unterscheidung zwischen 'seichtstruktur' und 'oberflächenstruktur' bei Cresswell (1973) oder in der unterscheidung zwischen PREAL MANIF und MANIF bei Grunig (1981) hat, erlaubt es, das instrument der kontextfreien

syntaxen erheblich effektiver zu nutzen, als dies gemeinhin getan wird.

Anm. 7: Betroffen sind alle varianten separierter syntaxen, für die grundsätzlich die substitution von minimaleinheiten durch komplexe ausgeschlossen ist (s. Grunig 1981: 603ff.), also z.b. syntaxen, wie sie im stil von GB, LFG oder GPSG konstruiert werden . - Knotenreichtum wird auch hervorgerufen durch einen hierarchischen aufbau der kodierung der kombinatorischen eigenschaften nach der art des sog. X-schemas.

Anm. 8: Der wert von τ ist sinnvollerweise abhängig zu machen von dem gewicht, das man dem jeweils in frage kommenden belegtyp geben will. Sei etwa eine syntax Σ empirisch zu rechtfertigen:

(a) $A \rightarrow B\ (C)$

(b) $B \rightarrow (D)\ E$

(c) $C \rightarrow F + B$

(d) $E \rightarrow (G)\ H$

(e) $H \rightarrow (I)\ J$

(f) $J \rightarrow (K)\ L$

(g) $L \rightarrow M + E$

Für die empirische rechtfertigung von B in (c) ist es offenkundig bedeutsamer, in den mit B zu beschreibenden ausdrucksstücken die anwesenheit von ausdrucksstücken zu belegen, die als D beschrieben werden sollen, denn die anwesenheit solcher stücke, die G, I, K zuzuordnen sind. Für die etablierung von B in (c) - anstelle etwa von E, H, J oder L etwa - sind die belege für D, G. I und K unterschiedlich wichtig; sie verlieren an bedeutung in der angegebenen reihenfolge. - So ist für die beschreibung der internen struktur von niederländischen ausdrücken, die mit *terwijl* eingeleitet werden, ein beleg, der zeigt, daß darin *zodat* auftreten kann, bedeutsamer als ein beleg mit *ook* oder *goede*.

(i) Door dit gelijke spel is Midstars nog steeds in de greep van degradatie, *terwijl* HIA-Panels nu gelijk is

gekomen met de Middenstandsbank, *zodat* de strijd HIA-Panels-Middenstandsbank wel eens beslissend kan worden.

(ii) Dat is vooral te danken aan het opgang komen van de activiteiten in de horecasector, terwijl het *goede* weer *ook* een rol gespeeld heeft.

Eine ausführlichere darlegung zur festsetzung der werte für τ findet sich bei Clément & Thümmel (1979: 113ff.).

Rieger, Burghard (ed.)
Glottometrika 13
Bochum: Brockmeyer 1992, 287-298

Das Problem der Datenhomogenität

Gabriel Altmann

1. In der angewandten Linguistik, wie z.B. in der Grammatikforschung, geht man davon aus, daß der Gegenstandsbereich völlig homogen ist (vgl. Altmann 1987). Auf diese Weise ist es überhaupt möglich, Regeln zu etablieren, Schulgrammatiken und Sprachlehrbücher zu schreiben. Was von dem homogenen Bild der Sprache abweicht, wird als dialektal, fachsprachlich, schichtspezifisch u.ä. eliminiert oder ignoriert.

Trachtet man aber nach dem Aufbau einer Sprachtheorie, dann muß man der Tatsache der Heterogenität Rechnung tragen. Da man homogene Daten in der Sprache selten findet, muß man die Heterogenität entweder in die Modelle einbauen oder die Daten so aufspalten, daß die resultierenden Klassen relativ homogen sind. Man muß nämlich bedenken, daß unsere Beobachtungen keineswegs Daten *sind*, sondern von uns zu Daten *gemacht werden*, und zwar im Lichte einer Theorie oder zumindest im Lichte eines Modells, in das entsprechende Annahmen aufgenommen wurden. Illustrieren wir einige Heterogenitäten an Beispielen:

(a) Frequenzwörterbücher, für die zahlreiche Texte verarbeitet wurden, stellen in jeder Hinsicht heterogene Daten dar. Sogar die Frequenz eines jeden einzelnen Wortes ist hier eine Mischung von heterogenen Frequenzen. Die Annahme, daß man sich einer Sprachnorm nähert, wenn man große Datenmengen erhebt, ist falsch. Häufigkeit ist sicherlich ein Attribut sprachlicher Einheiten, die das Köhlersche Anwendungsbedürfnis der Sprecher befriedigt (Köhler 1990), aber sie ist von so vielen lokalen (d.h. Autor-, Stil-, Textsorten- u.a.) Faktoren beeinflußt, daß es völlig illusorisch wäre, von einer "Häufigkeit des Wortes in der Sprache" zu sprechen. Nur die Eigenschaft "Häufigkeit des Wortes im gegebenen Text" reflektiert in idealen Fällen reelle Begebenheiten. Wenn man die Homogenität eines Frequenzwörterbuches annimmt, dann läßt man still-

schweigend die *ceteris paribus* Bedingung gelten, die sogar in kleinen Texten kaum erfüllt wird.

(b) Für eine Texteigenschaft, wie z.B. Satzlänge, ist nicht einmal ein einziger Text, z.B. ein Roman, homogen. Pausen beim Schreiben rufen womöglich Rhythmusveränderungen hervor, so daß man in dieser Hinsicht nur Textteile als homogen betrachten kann (vgl. Altmann 1988, 1988a: 68f.)

(c) Untersucht man die Abhängigkeit zweier Worteigenschaften voneinander, z.B. Länge und Polylexie (Polysemie), dann ist die Menge aller Wörter, ob nun aus Texten oder aus einem Lexikon, keine homogene Stichprobe, denn bei den Nomina kann die Abhängigkeit völlig anders aussehen als bei den Verben (vgl. Hammerl 1990).

(d) Die Häufigkeiten einzelner Bedeutungen, z.B. eines deutschen Nominalsuffixes, sind nicht unbedingt homogen, denn sie kann aus der Mischung der Häufigkeiten der einzelnen Wortarten, an die der Suffix angehängt wird, bestehen. Die Homogenisierung kann durch separate Untersuchung der einzelnen Wortarten erreicht werden (vgl. Altmann, Best, Kind 1987).

(e) Auch bei der Zählung der Phonem- oder Buchstabenhäufigkeiten unterliegt man dem Trugschluß, daß bei Millionen von gezählten Einheiten alle Inhomogenitäten beseitig werden. Im Gegenteil. Phoneme/Buchstaben kommen in Morphemen und Wörtern vor und hängen von deren Häufigkeit ab. Es gibt mehrere Morphem- und Wortarten mit völlig unterschiedlichen Phonem-/Buchstabenhäufigkeiten:

(i) Sehr häufig vorkommende Affixe einer Wortart im Lexikon, z.B. "-en" bei der Infinitivform deutscher Verben, unter der sie als Lexeme im Wörterbuch erfaßt werden. In einer Lexikonstichprobe erhöhen sie die Häufigkeit von "e" und "n".

(ii) Häufige Affixe besonders in synthetischen Sprachen, die man im Wörterbuch überhaupt nicht findet, z.B. Kasusendungen.

(iii) Hilfswörter, die es in allen Sprachen gibt, besonders in analytischen, z.B. Präpositionen, die in agglutinierenden Sprachen durch Affixe ersetzt werden.

(iv) Häufige Wörter des Grundwortschatzes, die man in allen Texten findet.

(v) Wörter, die man häufig in fachsprachlichen Texten, aber sonst selten findet.

(vi) "Normal" häufig vorkommende Wörter, z.B. in Prosa.

(vii) In Texten selten vorkommende Interjektionen, die oft besondere Phoneme enthalten.

Gilt eine Ranghäufigkeitsverteilung für Phoneme, dann gilt sie möglicherweise nur für eine dieser Klassen oder für jede separat (mit unterschiedlichen Parametern), aber nicht unbedingt für ihre Mischung. Es ist wahrscheinlich korrekter, die Verteilung der Phoneme nur z.B. für Hilfswörter aus einer Mischung von Texten zu untersuchen, als für alle Wortarten zusammen in einem einzigen Text. In Anbetracht der bisher ermittelten Datenmengen ist diese Erkenntnis recht enttäuschend. Will man trotzdem Modelle für die letztgenannten Daten aufstellen, dann müssen sie dieser Tatsache Rechnung tragen, d.h. es können keine "einfachen" Modelle sein, sondern möglicherweise Mischungen oder Zusammensetzungen von Verteilungen.

Zu diesem Umstand gesellt sich noch ein zweites Problem. Durch Vergrößerung der Stichprobe werden Inhomogenitäten nicht unbedingt verwischt, sie können sich sogar stärker ausprägen. Untersuchungen von großen Materialmengen, z.B. aller Werke eines Autors oder einer Stichprobe von einer Million Wörtern aus 500 Texten, können dann derartige Daten liefern, die man mit keinem Modell erfassen kann.

Betrachten wir als Beispiel zwei Briefe von Goethe, in denen Grotjahn (1982) die Wörtlänge gemessen hat. In Tabelle 1 (vgl. Altmann 1988a) findet man die Anpassung der negativen Binomialverteilung an diese Briefe. Addiert man aber die Häufigkeiten in beiden Briefen, so ergibt sich eine schlechtere Anpassung. Wie man sieht, ist der Parameter k in den beiden Briefen sehr unterschiedlich, so daß man von einer "Wortlängenverteilung in Goethes Briefen" als einer homogenen Tatsache nicht sprechen kann, und noch weniger über die "Wortlängenverteilung bei Goethe".

Tabelle 1

	Brief Nr. 612		Brief Nr. 647		612 + 647	
x	fx	NPx	fx	NPx	fx	NP
1	164	162.61	259	259.16	423	422.36
2	105	104.38	132	125.65	237	230.40
3	35	38.81	37	46.65	72	84.96
4	15	10.94	19	15.55	34	26.32
5	1	3.26	6	4.89	7	7.38
6	-	-	1	2.10	1	2.58
	k	6.312		1.882		2.842
	p	0.898		0.742		0.808
	x^2	3.47		3.91		5.39
	FG	2		3		3
	P	0.18		0.27		0.15

Je größer die Stichprobe, desto mehr Klassen muß man bilden, desto mehr muß man die Bedingungen, unter denen die Zähleinheiten vorkommen, beachten, um homogene Klassen zu bekommen. Tut man das nicht, dann sind die erhobenen "Daten" ungeeignet als Testinstanz für ein Modell, das von Homogenität ausgeht. Wahrscheinlich ist dies der Umstand, warum in der Linguistik vorläufig nur wenige Gesetze abgeleitet wurden.

2. Da unser Problem sehr allgemein ist und alle Bereiche der quantitativen Analyse betrifft, beschränken wir uns im folgenden auf Phonem-/Laut-/Buchstabenhäufigkeiten und werden nur von Phonemhäufigkeiten sprechen, was *mutatis mutandis* auch für Laute und Buchstaben gilt.

Phonemhäufigkeiten werden üblicherweise aus Texten oder einem Wörterbuch ermittelt, und es wird meistens stillschweigend angenommen, daß ihre Ranghäufigkeitsverteilung dem Zipf-Mandelbrotschen Gesetz folgt (vgl. Altmann 1988a; Zörnig, Altmann 1983, 1984). Dieses Verfahren, wie oben schon angedeutet, hat einige problematische Aspekte, die wir näher untersuchen wollen.

(a) Üblicherweise erhebt man große Stichproben, z.B. aus den Werken eines Autors, aus Texten eines Genres u.ä. Vergleicht man die Häufigkeit eines Phonems in zwei Teilen eines Textes, so läßt sich immer erreichen, daß auch der

kleinste Unterschied signifikant groß wird, wenn die Stichproben groß genug waren. Man kann auch die Umfänge berechnen, bei denen ein gegebener Unterschied signifikant wird. Ist dies aber der Fall, dann heißt es, daß die gegebenen Stichproben nicht zusammengefaßt werden dürfen. Daraus folgt weiterhin, daß man den Umfang so halten muß, daß der beobachtete Unterschied noch nicht signifikant groß wird. Dies ist natürlich nur ein Trick, mit dem man Homogenitäten erzeugen kann, aber der Wahrheitssuche ist er nicht dienlich. Es ist daher empfehlenswert, in einem Text die Phonemhäufigkeiten in seinen geschlossenen Teilen, z.B. in einzelnen Kapiteln, separat zu ermitteln. Danach muß festgestellt werden, ob die Phonemhäufigkeiten im ganzen Text homogen verteilt sind, beispielsweise mit einem Chiquadrat-Test. Wenn dies zutrifft, dann kann man den ganzen Text als eine Stichprobe betrachten.

Grundsätzlich inkorrekt ist es, eine Stichprobe aus vielen Texten zu erheben, beispielsweise aus Texten eines gegebenen Genres. Eine Mischung ist erst dann erlaubt, wenn alle Einzelstichproben homogen sind. Dies ist eher ein Glücksfall, der selten auftritt. Wenn zwei Stichproben in bezug auf eine andere Variable homogen sind, beispielsweise Satzlänge, dann brauchen sie in bezug auf Phonemhäufigkeiten nicht homogen zu sein. Homogenität in einer Variablen läßt sich also auf andere Variablen nicht übertragen.

(b) Eine Stichprobe aus dem Wörterbuch ist etwas völlig anderes als eine aus Texten. Hier gibt es keine Wiederholung häufiger Wörter, aber der Fall (e,i) aus 1 kann hier eine starke Verzerrung hervorrufen. Die Phoneme in den Affixen müssen als eine separate Schicht betrachtet und gezählt werden.

Im Köhlerschen Selbstregulationsschema (Köhler 1986) spielen diese zwei Häufigkeitsarten eine unterschiedliche Rolle. Während die Häufigkeit aus Texten das Anwendungsbedürfnis befriedigt, bedient die Häufigkeit aus dem Wörterbuch das Kodierungsbedürfnis der Sprachträger, denn die Wörterbuchhäufigkeiten zeigen die Rolle des Phonems beim Aufbau von Spracheinheiten und nicht bei seiner Verwendung. Diese Tatsache könnte sich im Köhlerschen Schema in Form von Entropie- oder Redundanzmaßen niederschlagen.

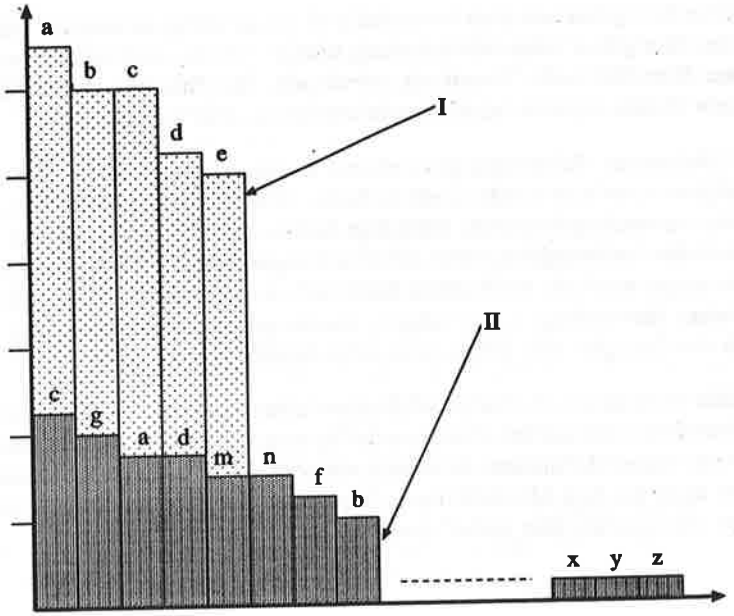
(c) Beim Modellieren der Ranghäufigkeitsverteilung muß man auf den Charakter der Stichprobe und auf den Charakter der Sprache achten. Die Häufigkeiten der Phoneme bilden hier mehrere Schichten, wie oben schon erwähnt. In bestimmten Fällen kann die Zipf-Mandelbrotsche Verteilung eine gute Anpassung liefern, theoretisch kann sie aber verfehlt sein.

Die Modelle der Ranghäufigkeitsverteilungen müßten also immer eine Überlagerung von Verteilungen darstellen, wenn man die einzelnen Schichten voneinander nicht trennt. Man kann sich die Situation folgendermaßen vorstellen. Seien Phoneme a, b, c, d, e bevorzugt in Affixen; ihre Häufigkeiten seien in Abbildung 1 mit der Verteilung I dargestellt. Die Häufigkeiten der außeraffixalen Phoneme seien mit der Verteilung II wiedergegeben.

Durch Addition der Häufigkeiten von a, b, c, d, e in den beiden Reihen werden möglicherweise Reihenfolgenveränderungen auftreten, mit Sicherheit verwischt sich aber der reguläre Abfall in den beiden Reihen. Die Situation wäre problemlos, wenn man die übereinander liegenden Häufigkeiten addieren könnte, d.h. $f(a_I) + f(c_{II})$, $f(b_I) + f(g_{II})$ usw., was natürlich keinen Sinn ergibt.

In einer derartigen Situation ist die gute Anpassung einer "glatten", monoton fallenden Verteilung an die Gesamthäufigkeiten als ein empirischer "Treffer" zu werten; von theoretischen Standpunkt verfehlt sie ihr Ziel, da sie nur eine grobe Approximation sein kann.

Abbildung 1



Eine mögliche Abhilfe in dieser Situation liefern uns die Überlagerungen von Verteilungen, von denen wir zuerst einige Spezialfälle erwähnen werden:

(1) Sei der Definitionsbereich der Ränge $x = 1, 2, \ldots n$. Die Verteilung bestehe aus zwei Komponenten. Die Variable der ersten Komponente mit dem Gewicht α ist definiert für $x = 1, 2, \ldots, m < n$, die der zweiten Komponente mit dem Gewicht $1 - \alpha$ ist definiert für $x = 1, 2, \ldots, n$. Die beiden Komponenten folgen der gleichen Verteilung mit identischen Parametern, die erste Komponente stellt eine rechts gestutzte Verteilung dar. Falls das ganze Modell durch Stutzung einer Verteilung mit dem Definitionsbereich $x = 1, 2, \ldots, \infty$ entstand, dann handelt es sich um die Mischung zweier Komponenten mit unterschiedlichen Rechtsstutzungen.

Beispiel. X folge einer rechts gestutzten Zipf-Mandelbrot-Verteilung mit

$$P_x = T\,(x+a)^{-b}\ ,\ x = 1, 2, \ldots, n\,;$$

wobei $a, b > 0$ und $T^{-1} = \sum_{j=0}^{n} (j+a)^{-b}$.

Für unsere Zwecke ergibt sich daraus die Mischung

$$P_x = \alpha\, T_1\,(x+a)^{-b} + (1-\alpha)\, T_2\,(x+a)^{-b}\ ,\ x = 1, 2, \ldots, n$$

wobei α $(0 < \alpha < 1)$ der Gewichtungsfaktor ist, und T_i ist definiert als

$$T_1 = \begin{cases} 0 & \text{für } x > m \\ \left(\sum_{j=1}^{m} (j+a)^{-b}\right)^{-1} & \text{für } x = 1, 2, \ldots, m \end{cases}$$

$$T_2^{-1} = \sum_{j=1}^{n} (j+a)^{-b}.$$

(2) Der Definitionsbereich der Ränge ist gegeben wie oben. Die erste Komponente mit dem Gewicht α ist definiert für $x = 1, 2, \ldots, n$, die zweite Komponente mit dem Gewicht $1-\alpha$ ist definiert für $x = c, c+1, c+2, \ldots, n$, wobei $c > 1$. Dieser Typ der Verteilung wurde von Dacey (1964a,b, 1966, 1970) eingeführt (wobei $x = 0, 1, 2, \ldots,$ und $c = 1$). Andere Mischungen dieser Art lassen sich durch

Entwicklung einer Verteilung aufstellen (vgl. Altham 1978; Gokhale 1975; Rutherford 1954; Shenton, Skees 1973).

Beispiel: Sei eine zweiseitig gestutzte Poisson-Verteilung (gestutzt im Punkt 0 und n) gegeben als

$$P_x = T\frac{a^x}{x!} \quad \text{für } x = 1, 2, ..., n$$

mit $a > 0$ und $T^{-1} = \sum_{j=1}^{n} \frac{a^j}{j!}$. Dann ergibt sich daraus

$$P_x = \alpha T_1 \frac{a^x}{x!} + (1-\alpha) T_2 \frac{a^{x-c+1}}{(x-c+1)}, \quad x = 1, 2, ..., n$$

wobei $0 < \alpha < 1$

$$T_1^{-1} = \sum_{j=1}^{n} \frac{a_j}{j!}$$

$$T_2 = \begin{cases} 0 & \text{für } x < c \\ \left(\sum_{j=c}^{n} \frac{a^j}{j!}\right)^{-1} & \text{für } x = c, c+1, ..., n \end{cases}$$

(3) Eine dritte Möglichkeit ergibt sich, indem man die Verteilung aus zwei Komponenten der gleichen Verteilung zusammensetzt, d.h. mit gleichem Definitionsbereich, jedoch mit unterschiedlichen Parametern. Je nach der Anzahl der Parameter ergeben sich hier immer mehrere Möglichkeiten.

Beispiel. X folge einer im Punkt 0 und Punkt n gestutzten negativen Binomialverteilung, gegeben als

$$P_x = T\binom{k+x-1}{x} q^x, \quad x = 1, 2, ..., n$$

wobei $k > 0$, $0 < q < 1$ und

$$T^{-1} = \sum_{j=1}^{n} \binom{k+j-1}{j} q^j$$

dann ist die Mischung gegeben als

$$P_x = \alpha T_1 \binom{k_1+x-1}{x} q_1^x + (1-\alpha) T_2 \binom{k_2+x-1}{x} q_2^x, \quad x = 1, 2, ..., n$$

mit $0 < \alpha < 1$,

$$T_1^{-1} = \sum_{j=1}^{n} \binom{k_1+j-1}{j} q_1^j \text{ und } T_2^{-1} = \sum_{j=1}^{n} \binom{k_2+j-1}{j} q_2^j,$$

woraus sich drei Möglichkeiten ergeben (i) $k_1 = k_2$, $q_1 \neq q_2$; (ii) $k_1 \neq k_2$, $q_1 = q_2$; (iii) $k_1 \neq k_2$, $q_1 \neq q_2$

Der allgemeine Fall, bei dem es mehrere Komponenten geben kann, die sich intervallweise überlagern, würde folgendermaßen aussehen:

Sei X definiert für $x = 1, 2, ..., n$, wobei in Punkten 0 und n im Modell eine Stutzung vorliegen kann. Für die Zahl der Komponenten (K) gilt die folgende Bedingung: Die Zahl der zu schätzenden Parameter im Modell soll kleiner als n-1 sein. Diese Bedingung ist besonders im Fall (3) zu beachten, wo die Zahl der Parameter schnell wächst.

Das allgemeine Modell für K Komponenten läßt sich dann schreiben als

$$P_x = \sum_{i=1}^{K} \alpha_i T_i P_{x_i} \quad x = 1, 2, ..., n$$

wobei $\alpha_1 + \alpha_2 + ... + \alpha_k = 1$, die i-te Komponente ist definiert für $x = L_i, L_i+1, ..., R_i$ und

$$T_i = \begin{cases} \left(\sum_{j=L_i}^{R_i} P_j \right)^{-1} & \text{für } x = L_i L_{i+1}, \ldots, R_i \\ 0 & \text{sonst,} \end{cases}$$

d.h. in der obigen Summe treten nur die Wahrscheinlichkeiten aus dem Intervall $[L_i,R_i]$ für die i-te Komponente auf. Die Intervalle einzelner Komponenten können sich ohne weiteres überschneiden, die Gewichtungen gewährleisten, daß $P_x = 1$ wird. Dabei kann eine Komponente auch im ganzen Definitionsbereich gelten, falls die Parameter der Komponenten identisch sind. Falls sie nicht identisch sind, dann können auch mehrere Komponenten im gesamten Bereich (1,...,n) definiert sein.

Die Mühsamkeit der Schätzung von vielen Parametern braucht nicht betont zu werden. In empirischen Wissenschaften findet man daher meistens nur Mischungen von zwei Komponenten.

Vom theoretischen Standpunkt ist dieses Verfahren korrekt, solange es sich nicht um Ranghäufigkeitsverteilungen handelt. Bei diesen kommt es durch Addition der Häufigkeiten identischer Einheiten des öfteren zur Unterbrechung von Intervallen, so daß der reelle Definitionsbereich einer Komponente aus mehreren unzusammenhängenden Intervallen bestehen kann. Vorläufig sehen wir keine Möglichkeit, dieses Problem theoretisch und praktisch zu bewältigen. Es bleibt also die einzige Möglichkeit, sich nach folgenden Regeln zu richten:

(1) Man trenne die für Zählungen gebrauchten Einheiten in so viele homogene Klassen, wie es nach der Eigenart der Sprache, des Textes und der Einheiten nötig ist.

(2) Bei Texten zähle man nur geschlossene Textteile aus, es sei denn, man führt die Zählung für (qualitativ) homogene Klassen durch. In dem Falle besteht die Möglichkeit, daß auch eine Zufallsstichprobe aus mehreren Texten (Textteilen) korrekt sein könnte. Dies hängt aber von der Beschaffenheit der betreffenden Einheit ab.

(3) Bei Zählungen in einem Lexikon führe man in jedem Fall die Trennung in homogene Klassen streng durch.

Literatur

Altham, P. (1978), "Two generalizations of the binomial distribution". *Applied Statistics 27*, 162-167.

Altmann, G. (1987), "The levels of linguistic investigation". *Theoretical Linguistics 14*, 227-239.

Altmann, G. (1988), "Verteilung der Satzlängen". *Glottometrika 9*, 147-170.

Altmann, G. (1988a), "Wiederholungen in Texten". Bochum: Brockmeyer.

Altmann, G., Best, K-H., Kind, B. (1987), "Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation". *Glottometrika 8*, 130-139.

Gokhale, D.V. (1975), "Indices and models for aggregation in spatial patterns". In: Patil, G.P., Kotz, S., Ord, J.K. (Eds.), *Statistical Distributions in Scientific Work*, Vol. 2. Dordrecht: Reidel, 343-353.

Grotjahn, R. (1982), "Ein statistisches Modell für die Verteilung der Wortlänge". *Zeitschrift für Sprachwissenschaft 1*, 44-75.

Hammerl, R.(1990), *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Bochum, Habilitationsschrift.

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

Köhler, R. (1990), "Elemente der synergetischen Linguistik". *Glottometrika 12*, 179-187.

Rutherford, R.S.G. (1954), "On a contagious distribution". *Annals of Mathematical Statistics 25*, 703-713.

Shenton, L.R., Skees, P. (1970), "Some statistical aspects of amounts and duration of rainfall". In: Patil, G.P. (Ed.), *Random counts in scientific work*, Vol. 3. University Park: The Pennsylvania State UP, 73-94.

Zörnig, P., Altmann, G.(1983), "The repeat-rate of phoneme frequencies and the Zipf-Mandelbrot law". *Glottometrika 5, 205-211.*

Zörnig, P., Altmann, G. (1984), "The entropy of phoneme frequencies and the Zipf-Mandelbrot law". *Glottometrika 6*, 41-47.

Rieger, Burghard (ed.)
Glottometrika 13
Bochum: Brockmeyer 1992, 299-300

Index et concordance pour «Alice's Adventures in Wonderland» de Lewis Carroll, by Philippe Thoiron and Alain Pavé

Travaux de linquistique quantitative, 16. Editions Slatkine, Paris & Geneva, 1989. xix + 119 pages + 10 microfiches.

Reviewed by Sheila Embleton, Toronto.

This volume is in effect a companion volume to Thoiron (1980). The various methodological problems (together with their practical solutions) encountered in the construction of the concordance and the indexes are discussed in some detail in that earlier volume. For a brief description of the contents of Thoiron (1980) together with a review, see Embleton (1984).

Microfiche technology has enabled the compact and economical dissemination of what would otherwise be a book of roughly 1000 pages more than the present one. The first part of the "hard-copy" itself here is more or less just a description of the various coding systems used and of the contents of the various microfiches, all with adequate exemplification to make everything quite clear. The larger part of the hard-copy is made up of a global index of lemmas arranged alphabetically (with frequencies given for the narrative, dialogue, and poetry portions of the text, as well as for the total work), the same information presented in order of decreasing frequency (for narrative, dialogue, poetry, and total), a summary table of the distribution of frequency of lemmas, and an index of grammatical categories (with summary tables by chapter and narrative/dialogue/poetry/total). The microfiche portion of the publication is made up of a chapter-by-chapter index of lemmas (2 fiches), a global index of forms (1 fiche), a chapter-by-chapter index of forms (2 fiches), and the concordance itself (5 fiches). Each fiche holds the equivalent of up to 98 hard-copy pages. For easy reference, each microfiche is labelled on top (in type visible to the naked eye) with its exact contents. Researchers unfamiliar with the use of microfiche need not fear - they are extremely readable, no less so than ordinary printed text, but of course one does need access to a microfiche reader, which may limit one's access marginally.

Thoiron (1980) is of the theoretical interest and tremendous practical use to anybody intending to construct a concordance of their own to any text. In addition, it serves as an example of the type of statistical study that can subsequently be performed, showing also how new hypotheses can be generated and fresh perspectives can be brought to bear on a text through the use of quantitative methods. The current volume is a very specialized and precise tool, and will therefore be of use to an entirely different audience, namely those who are specifically doing research on «Alice's Adventures in Wonderland» and who are in need of a concordance and index to it. Both volumes, although in entirely different ways, are extremely valuable contributions to their respective audiences.

REFERENCES

Embleton, Sheila (1984), "Review of Thoiron 1980". In: Rothe, U. (ed.), *Glottometrika* 7, p. 168-170. Bochum: Brockmeyer.

Thoiron, Philippe (1980), "Dynamisme du texte et stylostatistique: élaboration des index et de la concordance pour Alice's Adventures in Wonderland. Problèmes, méthodes, analyse statistique de quelques données". *Travaux de linguistique quantitative*, 11. Geneva: Slatkine.

Addresses

Prof. Dr. Gabriel Altmann, Ruhr-Universität Bochum, Sprachwissenschaftliches Institut, Postfach 10 1 40, 4630 Bochum 1, Germany

Prof. Dr. Moisei Boroda, Ruhr-Universität Bochum, Sprachwissenschaftliches Institut, Postfach 10 21 40, 4630 Bochum 1, Germany

Prof. Sheila Embleton, Dept. of Languages, Literatures and Linguistics York University, 4700 Keele Street, North York, Ontario M3J 1P3, Canada

Dr. Rüdiger Grotjahn, Ruhr-Universität Bochum, Seminar für Sprachlehrforschung, Postfach 10 21 48, 4630 Bochum 1, Germany

Jacques B.M. Guy, Telecom Research Laboratories, P.O. Box 249, Clayton 3168, Australia

Dr. sc. Viktor Krupa, KO SAV, Klemensova 27, Bratislava, CSFR

Tatsuo Miyajima, The National Language, Research Institute, 3-9-14 Nishigaoka, Kita-ku, Tokyo 115, Japan

Dr. Ursula Rothe, Ruhr-Universität Bochum, Sprachwissenschaftliches Institut, Postfach 10 21 40, 4630 Bochum 1, Germany

Prof. Dr. Wolf Thümmel, Waldemarstr. 2, 5600 Wuppertal, Germany

Dr. Joel Walters/Dr. Yuval Wolf, Bar-Ilan University, I-52900 Ramat-Gan, Israel

Peter Zörnig, Rohdesdik 45, 4600 Dortmund 15, Germany