THE LEGACY OF ZELLIG HARRIS I

AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE

Volume 228

Bruce E. Nevin (ed.)

The Legacy of Zellig Harris
Language and information into the 21st century
Volume 1: Philosophy of science, syntax and semantics

THE LEGACY OF ZELLIG HARRIS

LANGUAGE AND INFORMATION INTO THE 21ST CENTURY

VOLUME I: PHILOSOPHY OF SCIENCE, SYNTAX AND SEMANTICS

Edited by

BRUCE E. NEVIN
Cisco Systems, Inc.

JOHN BENJAMINS PUBLISHING COMPANY
AMSTERDAM/PHILADELPHIA

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences — Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data for volume 1

The Legacy of Zellig Harris: language and information into the 21st century / Edited by Bruce E. Nevin.
p. cm. -- (Amsterdam studies in the theory and history of linguistic science. Series IV, Current issues in linguistic theory, ISSN 0304-0763 ; v. 228)
Includes bibliographical references and index.
Contents: v. 1. Philosophy of science, syntax and semantics.
1. Harris, Zellig S. (Zellig Sabbettai), 1909- 2. Linguistics. I. Nevin, Bruce E. II. Series.
P85.H344 L44 2002
410--dc21 2002074704
ISBN 90 272 4736 6 (Eur.) / 1 58811 246 2 (US) (Vol. I)

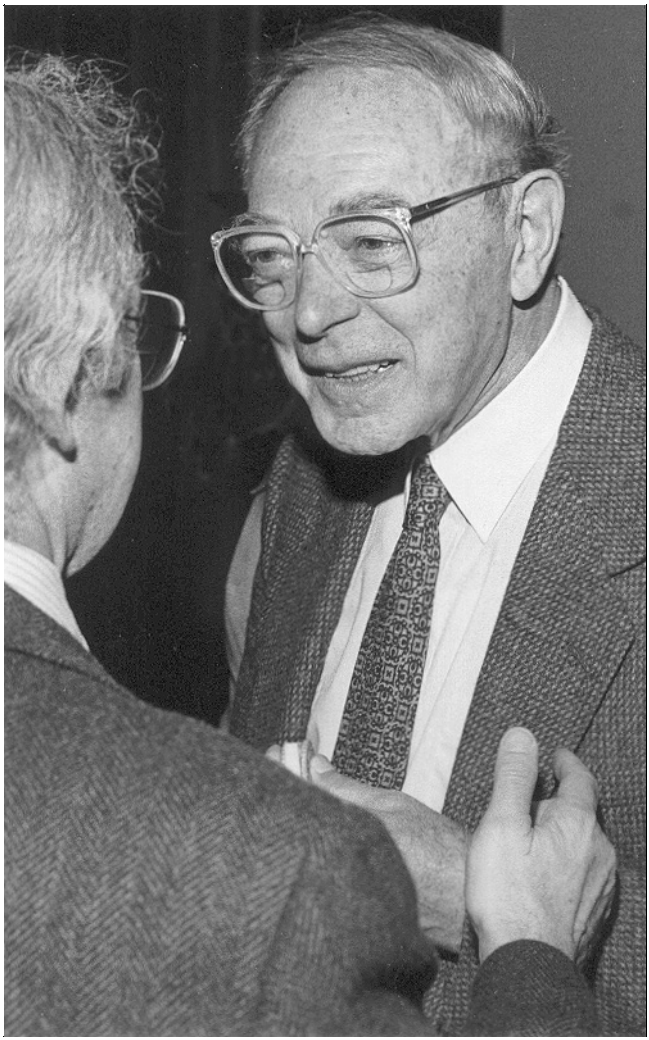
Volume 2

The Legacy of Zellig Harris: language and information into the 21st century / Edited by Bruce E. Nevin and Stephen M. Johnson.
p. cm. -- (Amsterdam studies in the theory and history of linguistic science. Series IV, Current issues in linguistic theory, ISSN 0304-0763 ; v. 229)
Includes bibliographical references and index.
Contents: v. 2 Mathematics and computability of language.
ISBN 90 272 4737 4 (Eur.) / 1 58811 247 0 (US) (Vol. II)

ISBN 90 272 4741 2 (Eur.) / 1 58811 316 7 (US) (SET VOLUME I+II)

John Benjamins Publishing Co. • P.O.Box 36224 • 1020 ME Amsterdam • The Netherlands
John Benjamins North America • P.O.Box 27519 • Philadelphia PA 19118-0519 • USA

Contents



Prof. Zellig Harris speaking to Prof. William Evan
(Photograph by Joe Pineiro, Columbia University)

Foreword

Bruce Nevin
Cisco Systems, Inc.

The contributions to this volume reflect the influence that Zellig Harris has had in syntax, semantics, mathematical linguistics, discourse analysis, informatics, philosophy (philosophy of science in particular), phonology, and poetics, and even, as we see here, in the design of computer user interfaces — not considering in this context his influence in the understanding of social and political change through discussions and writings relating to what was called the Frame of Reference for Social Change (F.O.R.) as epitomized in Harris (1997). His career spanned almost sixty years, and contributors include colleagues from several continents and students from the 1930s through the 1980s. However, this is not a Festschrift or a Memorial Volume, something to which Harris was adamantly opposed throughout his life. The only criterion for inclusion was that some relationship to Harris's work be clearly represented, and contributions critical of Harris's views were explicitly invited. As may be expected, many of those invited declined because of prior commitments during the time available for the project, or for other reasons. One demurral that may interest some readers is that of Harris's most well-known student, Noam Chomsky, of which more will be said presently.

To make evident the cohesiveness of the themes and issues represented in these volumes and their particular importance to any science of language, we first consider their origins. Then we survey the contents of this volume. Finally, we indicate equally germane lines of research that are not represented here, and that also call for further development. A comprehensive bibliography of Harris's writings, compiled by Konrad Koerner, appears at the end of this volume.

1. Origins

Harris's long career at the University of Pennsylvania began in the Department of Oriental Studies with a strikingly insightful Master's thesis in 1932 on the

accidental origin of the alphabet from the misinterpretation of Egyptian hieroglyphics by neighboring peoples as an acrophonic system — a fortuitous mistake because "writing systems do not otherwise naturally reach single sound segments" (Harris 1991:169).[1] This is of particular interest because it hinges on the phonemic contrasts which make language possible and on problems of the phonemic representations which make linguistics possible. Two years later, he submitted a descriptive grammar of the Phoenician language as his PhD dissertation in the same department (published in 1936 by the American Oriental Society). He published a number of other works in Semitic linguistics in the 1930s and as late as the grammatical sketch of Hebrew (Harris 1941a).

But even during this early stage he was working more generally in descriptive linguistics[2] and was deeply concerned with its methodology. Although the first indication in his bibliography of these broader concerns is the work with Carl Voegelin on Hidatsa (Lowie 1939), he was already developing the substitution grammar of (Harris 1946) and presumably was doing discourse analysis (published for Hidatsa in (Harris 1952a)), because as early as 1939 he was teaching his students about transformations[3] which by his own account arose out of the work on discourse analysis. As he says in the introductory essay of this volume, the close interrelationships and general character of these three aspects of language structure — discourse structure, transformations, and relations of substitutability yielding a grammar of constructions — were evident from the outset, though obviously not in detail. With these insights came the task of working them out, not as philosophical claims supported by episodic examples, but with full responsibility to the data of language. We turn now to the nature of that responsibility, and why it is important.

---

1. The 1932 thesis is summarized in (Harris 1933) and briefly recapitulated in (Harris 1991:168–171) with references to subsequent research. The *OED* defines 'acrophony' as "the use of what was originally a picture-symbol or hieroglyph of an object to represent phonetically the initial syllable or sound of the name of the object; e.g. employing the symbol of an *ox*, '*aleph*,' to represent the syllable or letter *a*."

2. The term 'structural', very much a term to conjure with in post-war social sciences, was substituted for 'descriptive' in the title of Harris (1951[1946]) by editors at the University of Chicago Press. With some amusement, Harris told me (in conversation in 1969) that he didn't remember whether they had asked him or not.

3. Lisker, in email of 1 March 2000 to Bruce Nevin.

1.1 Unavailability of a prior metalanguage

Already in the reviews of Gray's *Foundations of Language* and Trubetzkoy's *Grundzüge* (Harris 1940 and 1941b) we see the methodological principles that directed his research for the next 50 years. The most important of these is that fundamental limitation which he later articulated as the unavailability of any a priori metalanguage in which to define the objects and relations in a language. In the essay that introduces this volume, Harris says that

> since it is impossible to define the elementary entities and constraints of a language by recourse to its metalanguage (since the metalanguage is itself constructed from those entities by means of those constraints), it follows that the structure of language can be found only from the non-equiprobability of combinations of parts. This means that the description of a language is the description of contributory departures from equiprobability, and the least statement of such contributions (constraints) that is adequate to describe the sentences and discourses of the language is the most revealing.

In any other science, the structures and informational capacities of language are tacitly assumed as a given, insofar as these resources of language are freely used to describe the subject matter of the field, but for linguistics they *are* the subject matter. The consequence is that the work of linguistic analysis can be carried out only in respect to co-occurrence relations in the data of language — what had come to be called distributional analysis.

1.2 An illustration in phonology

Harris did not put this in terms of a 'metalanguage' in the 1930s and 1940s, but it is not difficult to see that his later formulation is merely a new way of stating the reason for limiting investigation "to questions of distribution, i.e. of the freedom of occurrence of portions of an utterance relatively to each other" (Harris 1951[1946]:5). One way to show this is by a brief excursus into Harris's discussion of phonology in those early years.

In the review of Trubetzkoy (Harris 1941b:707), he said that "phonemes are not absolute but relative [. . .] what is relevant in phonemics is only the contrast between one group of sounds and another." This was not a new conception. As de Saussure put it (quoted by Ryckman in this volume) phonemes are "wholly contrastive, relative and negative". To identify phonemes relative to one another, one begins with "differences between morphemes, which is precisely the criterion for determining phonemes" (Harris 1941a:

147). Native speakers judge whether two utterances contrast or are repetitions. A segmentation of utterances (if need be, this can be an arbitrary segmentation to start with) serves to locate the distinctions relative to one another within each utterance, and to associate them with phonetic descriptors. Then

> [. . .] for convenience, we [. . .] set up as our elements not the distinctions, but classes of segments so defined that the classes differ from each other by all the phonemic distinctions and by these only. [. . .] The classes, or phonemes, are thus a derived (but one-one) representation for the phonemic distinctions. (Harris 1951[1946]:35)

Subsequent distributional analysis refines the segments and relations among them while at each stage of re-analysis preserving that one-one or 'biunique' mapping to the fundamental data of contrast.

Continuing with the review of Trubetzkoy, and bearing in mind that phonemic segments are always in one-one correspondence to the phonemic contrasts that they represent, not to the phonetic descriptors or 'phones' with which they are associated:

> [. . .] it is pointless to mix phonetic and distributional contrasts. If phonemes which are phonetically similar are also similar in their distribution, that is a result which must be independently proved. For the crux of the matter is that phonetic and distributional contrasts are methodologically different, and that only distributional contrasts are relevant while phonetic contrasts are irrelevant.
>
> This becomes clear as soon as we consider what is the scientific operation of working out the phonemic pattern. For phonemes are in the first instance determined on the basis of distribution. Two positional variants may be considered one phoneme if they are in complementary distribution; never otherwise. In identical environment (distribution) two sounds are assigned to two phonemes if their difference distinguishes one morpheme from another; in complementary distribution this test cannot be applied. We see therefore that although the range of phonetic similarity of various occurrences of a phoneme is important, it is the criterion of distribution that determines whether a given sound is to be classified in one phoneme or another. And when, having fixed the phonemes, we come to compare them, we can do so only on the basis of the distributional criterion in terms of which they had been defined. As in any classificatory scheme, the distributional analysis is simply the unfolding of the criterion used for the original classification. If it yields a patterned arrangement of the phonemes, that is an interesting result for linguistic structure.
>
> On the other hand, the types and degrees of phonetic contrast (e.g. whether all the consonants come in voiced and unvoiced pairs) have nothing to do with the classification of the phonemes; hence they do not constitute a necessary patterning. (Harris 1941b:709–710)

What Harris rejects here is the notion advanced by Trubetzkoy and others that phonetic distinctions are relevant for determining phonemic contrasts, instead of being more or less useful for representing them in an organized way. Bloch (1953) attempted to derive phonemic contrasts from distributional analysis of phones rather than taking the contrasts as the primitive data. Like many linguists, he assumed that "the facts of pronunciation [are] the only data relevant to phonemic analysis" (Bloch 1941:283), and that each phonemic segment has a one-one or 'bi-unique' correspondence to a phonetic characteristic. Chomsky (1964) inveighed against this view.[4] These are, in different ways, attempts to define phonemes absolutely by locating them in a descriptive framework that is already defined prior to one's working on the given language. Such a framework is a set of statements as to what is possible for languages. These are, obviously, metalinguistic statements, therefore in the metalanguage for natural language.[5] But, as we have seen, such statements cannot be formulated and communicated other than in language.[6] The metalanguage makes use of the resources of language, so it cannot be presumed for purposes of determining what those resources are. Absent an external, prior metalanguage, definitions can only be relative.

This is not to say that there are no absolutes in phonology. However, these absolutes, the foundation data for linguistics, are not defined in articulatory or acoustic terms, nor functionally, but in terms of native speakers' perceptions of contrasts (distinctions) between utterances.[7] The relation of type to token

4. Halle (1954:335), Chomsky (1957:234), and Chomsky (1975:34, 91–93) refer to the pair test as a fundamental cornerstone of linguistics, which entails that perceptions of contrast are the primitive data of phonology. However, they reject the corollary "that only distributional contrasts are relevant while phonetic contrasts are irrelevant" (Harris 1941b:709), and the consequences that this has for the enterprise of a *phonetically* defined universal alphabet of *phonemically* distinctive features. For further discussion see Nevin (1995).

5. More carefully: a statement that is universal, as these purport to be, is a fortiori a statement about any given language and therefore in the metalanguage for that language.

6. Symbolic and mathematical representations depend upon language for their definitions and function as convenient graphical abbreviations that in any science are in practice routinely read out as sentences in oral presentations. Even mathematics depends upon the 'background vernacular' of language, although, as Borel (1928:160) notes, this is usually unnoticed.

7. To be sure, the study of phonetics is systematized in a theoretical framework, and in the interaction of speech physiology and acoustics there can be preferential 'affordances' for effecting the contrasts of languages (to borrow Gibson's term without buying with it his

has its root here, in the distinctions that language users make between different utterances, ignoring differences that make no difference when one utterance is a repetition of another. Edward Sapir, who regarded Harris as his intellectual heir,[8] brought this psychological reality to the fore, and "is remembered especially as one who emphasized the need to study what speakers know and believe (perhaps unconsciously) about their language, not simply what they do when they speak" (Anderson 2001:11).

The locations of the contrasts relative to one another within utterances, and their association with phonetic features of speech, are determined by various substitution tests, the most exacting form of which is the pair test (Harris 1951[1946]:32, 38–39, 1968:21–23). Distributional criteria enable us to derive "fewer and less restricted elements" (Harris 1951[1946]:59) that preserve this one-one correspondence to the distinctions, but the distinctions remain the fundamental data, and all the rest is a matter of working out a perspicuous physical representation for them. Harris never confused the representation with the reality.[9]

---

psychology), as in 'quantal' aspects of speech. But even if all languages *can* be represented phonemically by a universal alphabet of phonetically specified elements — be it said, the question of the "intrinsic content of features" raised in Chomsky & Halle (1968) remains unresolved today — the contrasts that those elements represent must be determined relative to each other for each language individually, and only then can they be specified using that alphabet. The principle gain of such an alphabet is the direct comparability of language structures, which interested Harris greatly, but for that purpose real differences of structure must not be suppressed by a demand for universality.

8. P.c. Victor Golla, May 1992, reporting conversation with Helen Sapir Larsen, who dated Harris in New Haven, reconfirmed in email of 18 Dec. 2001; also p.c., Regna Darnell, 31 Oct. 1997, reporting her own conversations with Sapir's children.

9. After the writings of the 1940s cited above, Harris said little more about phonology, probably because for linguistics beyond phonology it matters little how morphemes are represented so long as they are reliably distinct. Indeed, even a representation so notoriously imprecise and inconsistent as standard English orthography is routinely used for study of syntax and semantics, and suffices even for the stochastic procedure giving a preliminary segmentation of utterances into morphemes (Harris 1955). Of course, syntactic and semantic phenomena influence details of phonology pertaining to stress, intonation, and the like, and a few encapsulated relations such as rhyme and sound symbolism go the other way, but these are probably elaborations of pre-language modes of communication that are not unique to language, and in any case do not bear on the informational capacity of language, which was Harris's primary interest.

1.3 The role of linguistic intuitions

Another early statement of Harris's methodological stance is in the review of Gray's 1939 book (Harris 1940), where he says that intuitive apprehensions of *langue* cannot be used a priori in linguistics as means for identifying and accounting for regularities that may be observed in *parole*:

> With an apparatus of linguistic definitions, the work of linguistics is reducible, in the last analysis, to establishing correlations. [. . .] And *correlations between the occurrence of one form and that of other forms yield the whole of linguistic structure.* The fact that these correlations may be grouped into certain patterned regularities is of great interest for psychology; but to the pattern itself need not be attributed a metaphysical reality *in linguistics.* Gray speaks of three aspects of language [. . .], basing himself on the langue-parole dichotomy of de Saussure and many Continental linguists. This division, however, is misleading, in setting up two parallel levels of linguistics. 'Parole' is merely the physical events which we count as language, while 'langue' is the scientist's analysis and arrangements of them. The relation between the two is the same as between the world of physical events and the science of physics. The danger of using such undefined and intuitive criteria as pattern, symbol, and logical a prioris, is that *linguistics is precisely the one empirical field which may enable us to derive definitions of these intuitive fundamental relationships out of correlations of observable phenomena.* (Harris 1940: 704; emphasis added)

The patterning and structure found by the scientist in language may (and surely does) have a 'metaphysical reality' and significance outside of linguistics, but not within linguistics, on pain of question-begging.

The same considerations may be seen not just in phonology but also in Harris's treatment of all aspects of language, including of course syntax and semantics. Informant perceptions of contrast vs. repetition may be considered to be one kind of 'mental' data.[10] The second kind of 'mental' data in

10. The term is used here advisedly, given commonplace attributions of 'anti-mentalism' to Harris. "In practice, linguists take unnumbered short cuts and intuitive or heuristic guesses [but it is necessary then] to make sure that all the information [. . .] has been validly obtained" (Harris 1951[1946]: 1–2). These mentalist intuitions and guesses, which are notoriously unstable and difficult to control, must be validated. The requirement is "not so much to arrive at [plausible] conclusions, as to arrive at them from first principles, from a single method which is [. . .] essential because of the lack of an external metalanguage" (Harris 1991: 28–29 n.6). Note that this is not the truism that data must be validly obtained, but rather a concern that conclusions bear a valid relationship to those data. Some claims associated with the Universal Grammar hypothesis blur the distinction between data and conclusions, making this point difficult to recognize.

Harrisian methodology are informant judgements as to whether a given utterance is in the language or not — or more subtly to what degree or in what contexts it is acceptable as an utterance in the language. These two touchstones of 'psychological reality' identify the fundamental elements of language on the one hand and the universe of language data[11] on the other. All else is distributional analysis, whose aim is to identify constraints on the combinability of elements, the 'departures from equiprobability' of their combinations. Any seeming constraint that can be eliminated by a more perspicuous analysis is a property of inadequate analysis, not a property of language. Hence the endeavor, from the beginning, to wrestle the data into a representation whose elements are as unconstrained and freely combinable as possible. All that then remains are those essential constraints that together construct language, a 'least grammar'.

1.4 Complete coverage

Another cardinal point of Harris's methodology is that "data can be usefully analyzed only in respect to a specified whole structure of a language, and not as selected examples in support of episodic grammatical rules or processes" (Harris 1991:15). Harris did use isolated examples to clarify methodological issues in earlier writings, notably in Harris (1951[1946]), but claims about the structure of language were always grounded in broad coverage.[12]

2. Work represented in this volume

From the simplicity, directness, and mathematical elegance of the resulting theory of language arises the great fecundity of Harris's work in so many fields. Some of its many ramifications are reflected in the contributions to these volumes.

11. Generatively, in the use of graded acceptability as a criterion for transformation.

12. Assuming that student discussion focused around this ms. in the late 1940s, one could speculate that this is one reason that argument about examples and counterexamples may have appeared to the young Chomsky to be how one does linguistics.

2.1 Harris's survey article

A paper by Zellig Harris surveying the development of his own work introduces this volume.[13] In this survey article, Harris makes it clear that he began work on linguistic transformations quite early. In fact, he was talking about transformations in his seminars as early as the end of 1939 and certainly by the term ending June 1941.[14] Also very clear is the centrality of mathematics from the beginning, and the importance of simplicity (what he later termed a 'least grammar') for getting the machinery of analysis and presentation out of the way of our seeing transparently the correlation of form with information in language.

2.2 Philosophy of science

In the section on philosophy of science, T.A. Ryckman's "Method and Theory in Harris's Grammar of Information" discusses Harris's methodology and especially the role of the metalanguage. To be sure, he touches on misrepresentations found in many accounts of the recent history of linguistics, but the more important discussion concerns just how this methodology results in a grammar of information, in which is laid bare the intuitively obvious but hitherto elusive correlation of linguistic form with meaning. Information theory, it will be recalled, concerns only a statistical measure of the amount of information in a 'message' or through a 'channel'. It says nothing about information content.

Harris's theory of linguistic information is grounded in the same considerations of relative probability that underwrite mathematical information theory

---

13. This paper was first published as (Harris 1990) in French translation by Anne Daladier, who kindly provided the English text for this first publication in English, including several paragraphs toward the end that she received too late for translation and publication, which are published here for the first time. In one of these paragraphs, the opening sentence seems incomplete: "traditional and structural linguistics found regularities of co-occurrence word classes, between one or another word of another set." The author would surely have corrected this. The intention clearly is to say that traditional and structural linguistics dealt with selection restrictions only as far as word classes — more precisely, that they went only as far as saying that any word of one class could co-occur in a construction with any "one or another" of the words of a specified other class. The next sentences go on to say that it required first transformations and then operator grammar to "deal with the co-occurrence properties of individual words."

14. Lisker, in email of 1 March 2000 to Bruce Nevin.

as ordinarily understood. Given the social context of the origin, learning, and use of language, the commonplace hypothesis of a 'mentalese' for which language is a 'code', in the familiar conduit metaphor of communication, is superfluous. As a philosopher of science (concerned with questions posed by realist and instrumentalist interpretations of physical theories) and as a co-author of *The Form of Information in Science* (Harris et al. 1989), Ryckman is especially well suited to give this account of Harris's theory of information content.

Paul Mattick, another co-author of Harris et al. (1989), in his essay on "Some Implications of Zellig Harris's Work for the Philosophy of Science", shows the relevance of Harris's work to fundamental issues that are evoked in the philosophy of science by the question of a 'syntax of science'. Logical empiricism assumed that scientific thought could be adequately represented by a logical system. This had the advantage that meaning relationships and patterns of reasoning could be analyzed with some precision, but had little to do with the actual practice of science. More recent studies of science seek to specify boundaries of domains and the relations between methods, concepts, and data within them, but must depend for this on the much debatable semantic intuitions of the researcher. Harris's study of science sublanguages suggests an approach to the 'syntax of science' based not on philosophical presuppositions but the actual syntax of the specialized sublanguages used in doing science. It offers a middle way that combines analytical rigor with empirical sensitivity.

Maurice Gross turns directly to pivotal methodological (and empirical) issue that we have considered above, in his "Consequences of the metalanguage being included in the language". He clarifies the empirical basis of Harris's claim that the metalanguage of linguistics is included in each language that linguistics describes, and develops some theoretical implications which are perhaps surprisingly profound. These implications include the rejection of most categories currently in use in linguistics, such as subject, predicate, and semantic categories. When one takes the trouble to look, the generalizations expressed by these categories turn out to have as many exceptions as they have compliant examples. This gives striking support not only for their rejection, but also for a reconsideration of the notion 'grammatical exception'. Work on this paper was interrupted by the author's final illness, but although it might have been extended farther, it is complete in itself.[15]

---

15. The concluding sentence was added by the Editor (Bruce Nevin).

Noam Chomsky introduced the term 'discovery procedure' in 1957 in order to contrast it with the notion of 'evaluation procedure'. In "On Discovery Procedures", Francis Lin accepts the attribution of the idea of 'discovery procedures' to Harris, but under a reinterpretation in terms of language acquisition. Lin distinguishes two senses: mechanical procedures which the linguist follows to discover the grammar of a language, and innate procedures which enable the child to acquire the grammar of a language. Lin argues that Harris's formal procedures were not manual discovery procedures for the linguist[16] but can be regarded as innate procedures for grammar acquisition. He analyzes Chomsky's criticisms of Harris's 'discovery procedures' in both senses, and argues that the criticisms are unwarranted. He proposes that Harrisian constructions and transformations be used to explain knowledge of grammar and that Harrisian formal procedures be employed to account for grammar acquisition.

2.3 Discourse analysis and sublanguage grammar

The section on discourse analysis and sublanguage grammar begins with "Grammatical specification of scientific sublanguages" by Michael Gottfried, the first-named coauthor of Harris et al. (1989). The notion of sublanguage has an important place in Harris's perspective on language and its development. In Harris et al. (1989), this concept was extended to characterize the discourses concerned with a specific research question in immunology — that is, the sublanguage of that particular scientific domain. Harris et al. (1989) describes the grammatical organization of these discourses and the ways in which the information in them could be formally represented. Distinguished parts of these discourses were described in terms of (1) metascience operators such as *We hypothesize*, (2) science language sentences, and (3) various conjunctional and relational operators. Gottfried addresses some further questions involved in grammatical specification of a scientific sublanguage, in particular, the relation that the metascience operators bear to science language sentences. A sublanguage is defined by closure under operations of the grammar. New results presented here indicate that a science sublanguage is closed under resolution of referential forms.

---

16. For additional discussion of the observation that Harris did not intend his methods as practical discovery procedures, see e.g. Hymes & Fought (1981:176), Ryckman (1986: 45–54), Nevin (1993:375–376).

James Munz, in "Classifiers and Reference," identifies complications to the resolution of reference that may arise when the corpus for sublanguage analysis is less constrained than that of Harris et al. (1989). He reports how in a study of about 50 articles on the function of cardiac glycosides (the digitalis family) it was found that classifier vocabulary played a prominent role in discourse coherence. However, the cross-classification relationships turned out to be rather complex and in some measure ad hoc, and especially subject to change through time. These lexical relationships could be included in the grammar at the cost of complicating it. Munz discusses the tradeoffs in sublanguage grammar between a simplification of grammar at the cost of more complex treatment of pragmatics (the interpretation), or inclusion of such classifier relations at the cost of complicating the grammar and lexicon. He demonstrates these issues for a sublanguage of pharmacology.

Many linguists and rhetoricians have applied the term 'discourse analysis' to what might be called the macro-syntax of text, involving overt markers for such things as point of view, shift of topic, and conversational turn-taking, rather than with the information content of texts. In "Some Implications of Zellig Harris's Discourse Analysis", Robert Longacre explores the relationship of these approaches to that of Harris, using for demonstration purposes the "Proper Gander" text whose analysis Harris published early in the work on discourse analysis.

In "Accounting for Subjectivity (Point of View)", Carlota Smith develops what she calls an information-based approach to certain aspects of meaning in which a recipient constructs a point of view on a text by ascribing perspective, consciousness, an attitude, belief, or way of thinking to the author or to a person presented in the text on the basis of a composite of grammatical and lexical forms rather than any single feature. Smith has identified a number of principles of composition. There may be a single, sustained point of view; or more than one, conveyed by forms with a deictic component, including pronouns and anaphors, tense, adverbials, aspectual viewpoint, and modality. A clear point of view often emerges when terms from more than one deictic sub-system are in accord, and there may be rather subtle interactions between forms. For instance, whether or not tense is deictic depends to some extent on modality. In a sense this project relates directly to Harris's important work on co-occurrence, though Smith's interest is in the co-occurrence of what she terms linguistic features rather than of words or word-classes, and with the interpretations which recipients make of them.

2.4 Syntax and semantics

In much of the literature of linguistics, syntax and semantics are distinct rubrics. For Harris, they are two faces of the same socially maintained phenomenon, linguistic information, and this is borne out in the section on syntax and semantics. Harris was always interested in comparing different languages, e.g., "if such description is made of two or more languages it becomes relatively easy to compare their structures and see wherein and how far they differ" (Harris 1941a: 143).[17] Harris (1954) described how to create a transfer grammar expressing the relationship between the structures of two languages. Morris Salkoff is probably the first actually to write a detailed transfer grammar. In "Some New Results on Transfer Grammar", Salkoff sketches the syntactic portion of a French-English transfer grammar (Salkoff 1999), and describes the method by which it was developed. In the contrastive French-English grammar, the comparisons between French structures and their English equivalents are formulated as rules which associate a French schema (of a particular grammatical structure) with its translation into an equivalent English schema. For each of the principal grammatical structures of French — the verb phrase, the noun phrase and the adjuncts (modifiers) — the grammar presents all the rules that generate the corresponding English structure under translation. In addition to its intrinsic linguistic interest, this comparative grammar has two important applications. The translation equivalences that it contains can provide a firm foundation for the teaching of the techniques of translation. Furthermore, such a comparative grammar is a necessary preliminary to any program of machine translation, which needs a set of formal rules, like those given here for the French-to-English case, for translating into a target language the syntactic structures encountered in the source language. Of particular interest are ambiguities and word classifications that arise cross-linguistically but not in either language by itself.

In "Pseudoarguments and pseudocomplements", Pieter Seuren investigates a phenomenon that has been neglected in much of the linguistic litera-

17. Also: "in many [. . .] published grammars, the reader who wishes to have a picture of the language structure has to reanalyze the whole material for himself, taking merely the elements and their distributions as reported by the author of the grammar and hardly utilizing the author's analysis. This situation arises from the fact that linguists do not yet have a complete common language" (Harris 1944a: 190), i.e. a common metalanguage for the science of linguistics. This was one motivation for clarifying methodology.

ture.[18] Embedded infinitival, participial, adjectival, and prepositional clauses often occur as quasi-object-complements to main verbs whose meaning does not appear to call for a sentential object-complement. Such 'pseudocomplements' normally have a resultative, purposive, or comitative meaning. The English verb *go*, for example, is normally intransitive and its meaning does not call for a resultative, purposive or comitative complement proposition. Yet English has sentences like *John went fishing*, where *fishing* has the same syntactic features as it has in *John went on fishing*, where *fishing* is a real, semantically motivated object-complement. This phenomenon is extremely general, perhaps even universal. In almost all such cases, the original lexical meaning of the embedding predicate is 'bleached' and the verb in question assumes auxiliary status. Seuren hypothesizes that pseudocomplementation with *be* is the origin of the English progressive form.[19] In many languages, the distribution of infinitival and participial pseudocomplements (PCs) is restricted by the lexical specification of the embedding main verb. In SVC languages (with serial verb constructions), the distribution with regard to the embedding main verb appears to be free, but the selection of the embedded main verb is lexically restricted. Adjectival and prepositional PCs in sentences like *He painted the door green* or *He put the ladder up*, also extremely widespread in the languages of the world, appear to be adequately described as cases of pseudocomplementation. For this analysis to hold, it must be assumed, with Harris and McCawley, that all lexical content words, including adjectives and prepositions (those not used as argument indicators), are semantic predicates and thus take arguments.

In "Verbs of a feather flock together", Lila Gleitman reports on experimental work showing how structural (syntactic) information supports the learning of the semantics of verbs. Harris saw that a fine analysis of verbs'

---

18. It is related to the notion of *support verb* elaborated by Maurice Gross in his (1975) and in later developments of lexicon-grammar. Gross attributes its origin to Harris (1951[1946]).
19. Harris (1982:158) derives the progressive from a zeroed preposition: *John is fishing* ← *John is on his fishing*. Compare obsolete and dialectal *John is a-fishing*, where "in Old and Middle English *a-* (a reduced form of *on*) plus the verbal noun ['gerund' with *-unge*, *-ing*], was used for a second argument of *be*, *go*: *He was a-fighting*, *I go a-fishing*" (Harris 1982:295). It is often the case in operator grammar that zeroing of a higher operator (here, the preposition *on*) results in apparent multiple classification of the operator under it (of which the distinction between intransitive *go* and this 'pseudocomplement'-taking *go* is an example).

relative distribution with regard to their complement structures (their subcategorization privileges) effects a coarse semantic partitioning of verbs in the lexicon. Experimental work has shown, in various languages, how syntactic overlap predicts semantic overlap. For example, languages largely agree in reserving tensed sentence complements for verbs with mental content. Such correlations can be used to project the syntactic privileges of known verbs beyond their observed structural environments, and to project the meanings of verbs based on their observed structural environments. Studies of infants and toddlers by Gleitman and colleagues demonstrate that they use these regularities as guides to the interpretation of verbs. Such a structure-sensitive learning procedure is required in the vocabulary learning process because raw observation of the extralinguistic environments for verb use provides almost no constraints that can be used to derive their meanings. The use of structural evidence by children begins to explain the robustness of their vocabulary learning irrespective of vast differences in the extralinguistic evidence made available to them, as for example (dramatically) in the congenitally blind child's understanding of such 'visual' terms as *look* and *green*.

2.5 Phonology

Two papers here revert to the issues of phonetics and phonology which we used earlier to illustrate Harris's methodological principles. In "The Voiceless Unaspirated Stops of English", Leigh Lisker applies substitution tests not for identifying contrast vs. repetition, but for discriminating phonetically precisely where an identified contrast is located in the phonetic data of speech. The distinction between English /p/ and /b/ has generally been assumed to be effected by differences in voice onset time (VOT) during or after stop closure. Lisker demonstrates that this is not always the case, with evidence that may challenge some familiar and supposedly universal categories such as 'voiceless unaspirated'. For Harris, of course, this would not matter: the contrasts between the original utterances are already determined by the judgements of native speakers, and the substitution experiments serve only to locate the contrasts relative to one another and to associate them with phonetic data. Whatever the phonetic facts turn out to be, those they are. An a priori phonological framework that posits a universal feature expressing VOT such as [+delayed release] is problematic. This is not a case of neutralization, since in context the distinction is perceived, but whereas in some contexts the /p/ vs. /b/ distinction is effected by VOT, in these cases it

is effected by other means still to be determined among phonetic features preceding stop closure.

Sometimes what seems most obvious is in fact not well defined. Such is the familiar distinction between consonants and vowels, which is vexed by intersecting but non-identical phonetic and phonological usage of these terms and the persistent but still unachieved objective in Generative phonology of giving phonetic definitions for a universal set of distinctive, that is, phonemic, features. Harary & Paper (1957) described phoneme co-occurrence in abstract relational and graph-theoretical terms and suggested a numerical method, related to that of Harris (1955), for describing the distribution of the phonemes in a language. In "On the Bipartite Distribution of Phonemes", Frank Harary and Stephen Helmreich extend this work, developing the notion of a bipartite graph and applying it to the distribution of phonemes. Recognizing the intimate relation of the C–V distinction to the no less elusive notion of syllabicity, they propose a method for distinguishing *+syllabic* segments from *−syllabic* segments. First, they look at this distinction from the standpoint of graph theoretic concepts, and, building on the work of Harary & Paper (1957), they relate it to the graph-theoretic notion of a bipartite graph. In a bipartite graph *G* every node can be colored with one of two colors so that each edge of *G* joins two nodes of different colors. In this case, the 'colors' are vowel and consonant. They study the graph in which the nodes are phonemes and the edges are determined by succession in the corpus. They introduce a method of quantifying the degree of bipartiteness of the phonemic graph of a particular corpus, apply this method to Hawaiian, and show that the C–V division produces a highly bipartite graph. Second, they generalize this result by developing a computer program which examines divisions of elements of a set (in this case, phonemes) into two groups, including such phonetic divisions as front/back, high/low, etc. This program determines the bipartiteness of each division with respect to the graph of a corpus in that language. They show for a number of typologically diverse languages that the most bipartite graph has an obvious interpretation as the C–V distinction for that language, so that this approach provides a distributional method of identifying this important distinction.[20]

---

20. It will be interesting to see this method applied to languages in which segments that are consonants by any phonetic reckoning are syllabic, such as Imdlawn Tashlhiyt Berber (Prince & Smolensky 1993 and works of Dell and Elmedlaoui cited there).

2.6 Applications

Daythal Kendall presents an application of Harrisian linguistics to poetics in "Operator Grammar and the Poetic Form of Takelma Texts". Within the framework of Harris's operator grammar, Takelma texts turn out to be highly structured. The overall form of a fable, as revealed by syntactic analysis and other factors such as deviations from the dominant subject-object-verb (SOV) word order, repetitions, and overt lexical markers, indicates a deliberate manipulation of the language to create an artistic form. First, a partial syntactic analysis of the fable "Coyote's Rock Grandson" illustrates the technique, and then the entire fable is given in free-verse form, as recovered through syntactic analysis, in both Takelma and English. Finally, Kendall discusses symbols used in the story, social norms addressed, and details of the poetic form. In addition to providing entertainment, many Takelma fables illustrate the consequences of violating selected social conventions. "Coyote's Rock Grandson" addresses the issues of obtaining a wife, establishing and maintaining the proper social and economic relationships between a man and his wife's parents, and behaving appropriately toward one's neighbors. Explication of the text shows how the social fabric is rent and restored, and the consequences of violating conventions. In addition, Kendall discusses the use of symbols, the manipulation of syntax, overt surface markers, repetition, and the poetic form of the fable.

It is not often remembered today how closely the development of departments of linguistics in the United States was related to the development of new methods of teaching languages during and after World War II. From at least 1939, Zellig Harris taught a course called Linguistic Analysis in the Department of Oriental Studies and later in Anthropology. In 1946, a Department of Linguistic Analysis was established at the University of Pennsylvania, which was renamed the Department of Linguistics in 1948. Many of the early papers, such as those on Swahili, Moroccan Arabic, and Fanti, came out of this work. In a fascinating reprise of these concerns, still of vital importance today, Fred Lukoff gives us "A Practical Application of String Analysis" to language pedagogy, using the teaching of Korean as an example. Students' perception of the structure of long Korean sentences can be hindered by the length and complexity of adjunct material (modifiers). Lukoff shows that long sentences in written Korean are well suited for analysis by the methods of Harris (1962), because the relatively fixed word order of Korean sentences allows comparatively straightforward progressive excising of adjunct material, leaving the minimal structure, or elementary sentence, which lies at the syntactic and

semantic core of the given sentence. These methods lead to an uncovering of the basic constituent structure and hence to an understanding of the intended meaning of long sentences in Korean. Lukoff draws numerous examples from texts of various kinds, and gives step-by-step string analysis procedures for uncovering the structure of these sentences. Classroom experience with string analysis has proven to be an effective and efficient means by which students learn to perceive the structure of long, complex Korean sentences. He gives examples of the difficulties that students typically experience when they lack these tools. Applying the methods of string analysis also brings certain questions into focus, some new and some old, concerning the constituent structure of Korean sentences, among them the problem of analysis of ambiguous coordinate constructions. There are obvious and interesting relations between this paper and that of Salkoff on transfer grammar. I am happy to report that Professor Lukoff was able to complete this work just before his death, and that he was pleased with the result.

3. Further issues and themes

Developments based on Harris's work in mathematical linguistics, formal systems, informatics, and computer science are presented in Volume 2 of this work. Many other researchers who were invited were unfortunately unable to prepare a contribution within the time available for this project, or declined for other reasons. There are also themes and problems for research, many of which Harris identified, that are not reported in this publication.

3.1 Relation to Generativist theory

Noam Chomsky was invited to contribute to these volumes on any topic, at any length, with particular interest being expressed in his perception of the relation of his work to that of Harris. He declined on grounds of having too slight acquaintance with Harris's work after the mid 1950s.[21] It is unfortunate

21. More specifically, that he did not fully understand what Harris was doing and why after the mid 1950s because, in his view, Harris's work then took a new course, which he had not followed closely. Compare Harris's account, in the present volume, of continuous development of principles evident from the beginning. As to Chomsky's understanding of those principles and of Harris's earlier work, a comparison of their respective accounts is

that his remarks in various places generically about 'structural' or 'taxonomic' linguistics have frequently been understood as applying to Harris; they do not.[22]

Chomsky's most substantive contributions — and they are very important and lasting — are in the characterization of formal systems, such as the Chomsky hierarchy of context-free languages. For Harris, rather

> The interest [. . .] is not in investigating a mathematically definable system which has some relation to language, as being a generalization or a subset of it, but in formulating as a mathematical system all the properties and relations necessary and sufficient for the whole of natural language. (Harris 1968:1)

Symbolic approaches such as phrase-structure-grammar (PSG) build abstract structures, typically represented as labeled trees. Relations between words are mediated by abstract preterminal nodes that are devoid of semantic content. Lexical content is inserted into an abstract structure, and special mechanisms are adduced to account for meanings and selection restrictions on the one hand and for phonological phenomena on the other. Each such mechanism has a distinct syntax and semantics over its own vocabulary of (typically) features and feature values. In any Harrisian grammar, from the 1940s to 1991, sentences are constructed out of words directly, and selection restrictions and phonological changes apply at a stage of construction when the relevant lexical items are contiguous.

> There is an advance in generality as one proceeds through the successive stages of analysis [from structural linguistics, to transformational analysis, to operator grammar]. This does not mean increasingly abstract constructs; *generality is not the same thing as abstraction*. Rather, it means that the relation of a sentence to its parts is stated, for all sentences, in terms of fewer classes of parts and necessarily at the same time fewer ways ('rules') of combining the parts, i.e. fewer constraints on free combinability (roughly, on randomness). But at all stages the analysis of a sentence is in terms of its relation to its parts – words and word sequences – without intervening constructs. (Harris 1981: v; emphasis added)

---

easy for anyone to make, though it is frequently unclear which if any characteristics attributed to structural or 'taxonomic' linguistics are intended to apply to Harris's work. Chomsky (1977b:122), in a paragraph that was inserted into the translation of Chomsky (1977a), mentions Harris (1965), but probably only referring to Harris's objection (in fn. 5) to the pitting of one tool of analysis against another.

22. See Ryckman in this volume. Also Nevin (1995), a severely cut version of a ms. distributed at ICHoLS VI and revised slightly in 1999 in response to correspondence with Noam Chomsky in 1995 and 1997–1998.

It may be that the central difference between Generativist theory and Harris's theory of language and information is that the former asserts the existence of a biologically innate metalanguage.[23] The fact that grammars and a theory of language and information have been achieved without this hypothesis presents a considerable challenge to it, a kind of existence proof. Against this press the intellectual and professional commitments of a great many people constituting the field of linguistics as it stands today. In consequence, the question continues widely unaddressed.

Harris recognized no standpoint outside of language from which one may describe language. Without an understanding of this fundamental fact, the methodological strictures that Harris imposed on his work are incomprehensible. Those who (without understanding this) have attempted to account for the incomprehensible have described Harris as anti-mentalist, positivist, even behaviorist — preposterous falsehoods, trivially refuted by simple examination of his writings.[24] He did not even *like* the word 'behavior'.[25]

Some have asked why Harris did not defend himself against attack and refute at least the most ludicrously false attributions. He was certainly capable of intellectual criticism of great power and incisiveness where he felt that was necessary and appropriate, as in politics. For science, however, where unfettered inquiry and open communication are essential, scientific results must stand or fall on their merits, and need not, indeed should not be defended. He was dedicated to science and in particular to the methods that became clear in the late 1930s and early 1940s as ineluctable for any science of language. In any field — this is strikingly illustrated by his posthumous book on politics (Harris 1997) — he always looked for constructive solutions in a positive way. An

---

23. To be sure, Chomsky does not accept that UG constitutes or incorporates a prior metalanguage, but it can scarcely be disputed that it is given a priori, nor that it is metalinguistic (else how could it constrain grammars and languages).

24. These allegations are frequently made about 'structural linguists' generically. Chomsky (1964) provides an early example and Anderson (2001) a recent one. Whether or not these allegations are meant to apply to Harris is often either equivocal or not obvious. For example, we are to infer that Chomsky did not intend the 'Linearity Condition' to apply to Harris from his mention in a footnote (Chomsky 1964: 82n.16) that there are several examples in *Methods* (vaguely identified as in "chapters 7, 9") that pose difficulties for such a principle. See Nevin (1995) for discussion.

25. P.c., 1969 or 1970. Harris was talking with someone who had evidently called his attention to some statement or suggestion that he was a behaviorist. This was spoken aside to me, with something between amusement and bemusement at the preposterousness of it.

attack on the work of others benefits no one, and the perspicuity of an analysis is established not by argument but by demonstration in respect to the data of the field. Furthermore, he did not cling to his ideas and insights, however major, to defend them as his own, but instead with great creativity moved on to more advanced discoveries.

It has been alleged (following Chomsky) that Harris denied the reality of language and saw no value in asking how it is that language is uniquely learned by human children. Rather, he denied that philosophical speculation about these matters had any place as guiding principles for doing linguistics.[26] They can only fruitfully be asked in light of what linguistics has to tell us, since "linguistics is precisely the one empirical field which may enable us to derive definitions of these intuitive fundamental relationships out of correlations of observable phenomena" (Harris 1940:704). It was out of profound respect precisely for the *reality* of language that he would not interpose between linguists and the object of their investigation any a priori speculations about the 'patterned regularities' that may be observed in language. Harris employed the only kind of methods that can be available given the absence of any presupposed metalanguage.[27] As these methods disclosed more clearly the structure of language as a whole, and especially as operator grammar began to emerge in the late 1960s, he included in each report some statement of the interpretation of the formal results. The clearest and most complete statement is in his last book, *A Theory of Language and Information* (1991). Here, he dis-

26. No "metaphysical reality in linguistics" (Harris 1940:704, quoted earlier). For example:

> At various points, the conclusions which are reached here turn out to be similar to well-known views of language. [. . .] The intent of this work, however, was not so much to arrive at such conclusions, as to arrive at them from first principles, from a single method which is proposed here as being essential because of the lack of an external metalanguage. The issue was not so much what was so, as whether and how the essential properties of language made it so. (Harris 1991:28–29 n.6)

> That the picture of language presented here is similar in some respects or others to various views of language [. . .] does not obviate the need for reaching the present theory independently by detailed analyses of language. This is so partly because a more precise and detailed theory is then obtained, but chiefly because we then have a responsible theory in direct correspondence with the data of language rather than a model which is speculative no matter how plausible. (Harris 1991:6n.1)

27. Contrast e.g. "on suppose donnée à l'avance la structure générale d'un système de savoir" (Chomsky 1977a:126) or in the expanded English revision "one assumes the general form of the resulting system of knowledge to be given in advance" (Chomsky 1977b:117).

cusses how words carry meaning; what linguistic information is and how it is created in the form of predication in sentences; how language 'carries' information; the relation of form and content; the relation of language to the perceived world; the structure of information; the origin of language in prehistory; the nature of language as a self-organizing system in a context of social use; its relation to thought; and what capacities are required to learn and sustain language. It was never the case that Harris thought these questions were improper, only that asking them prematurely was, and that any answers to them that one might propose (necessarily a posteriori) have no a priori relevance among the data or 'facts' of linguistics. His methodological strictures can be seen as an injunction against question-begging. You have to know what you're talking about before you can talk about it. You can't get to language by way of talk about language, only by way of a nonverbal analysis of the objects and relations observed in language.

One may ask whether all of Harris's methodological strictures are still necessary for the description of languages. Perhaps some may now be set aside if they were required only for the achievement of a theory of language determined by "how the essential properties of language made it so" (Harris 1991:29). With that goal attained, perhaps we can now presume henceforth the minimalist characterization of language and information, the 'least grammar', that his methods at last reached.[28] We might take his theory of language and information to constitute at least the lineaments of a 'prior metalanguage', given now a priori in the science of linguistics, by which we may anticipate the structure of a given language that we aim to describe — a formulation, in short, of Universal Grammar.

But perhaps such an assumption carries no more weight than the "short cuts and intuitive or heuristic guesses" that linguists have always made. It is true that in addition to these 'intuitions of linguistic form' we now have an informed theory of language, but if the linguist fails to verify that conclusions about a particular language are validly related to the data of that language, it is at risk of having

28. Something like this is seen in Harris et al. (1989, Chap. 7) where the information formulae of a science sublanguage, established for English, are subsequently used to recognize the corresponding sublanguage categories and dependencies in French texts in the same domain. However, sublanguage analysis can avail itself of an external, a priori metalanguage without prejudice to the question we are considering here.

> the undesirable effect of forcing all languages to fit a single Procrustean bed, and of hiding their differences by imposing on all of them alike a single set of logical categories. If such categories were applied [. . .] it would be easy to extract parallel results from no matter how divergent forms of speech. (1951[1946]:2)

To this earlier caution, Harris might now reply that on the other hand anything that did not have the essential properties found for language — phonemic contrasts, constraints on combinability, especially dependence on dependence, and so on — would be something other than language. A pertinent question, then, is how much of this is in fact innate and how much is merely indispensable, given the informational uses of language.

3.2 Continuing themes

A number of themes cut across the topical sections into which the contributions have been organized in this volume, and bridge even across the two volumes. For example, though Harris did not propose an account of the development of language until late in his career (Harris 1988, 1991), his procedures for the linguist have always seemed to invite reinterpretation as processes for the child (e.g. Chomsky 1977a:124–128, English version revised and expanded in 1977b:114–119). This theme is explicit in the papers by Gleitman and by Lin, and in the paper by Pereira in Volume 2. Another recurrent theme is the role of classifier relations in vocabulary. Although these relations are notoriously inconsistent and unstable (see Munz in this volume), and for that reason Harris did not find them useful in grammar or in a theory of language, they may align with the lexical classes in sublanguage grammar where they are defined in an explicit external metalanguage, and they may further be useful in resolution of referentials. Papers here by Mattick, Munz, Gottfried, and Gleitman touch on these issues, as does that by van den Eynde et al. in Volume 2. The attentive reader will discern other shared themes. Broached in a number of places but not yet frankly advanced here is the very important question of argument and consequence in science,[29] one of a dozen or so topics called out by Harris (1991:28) as suggestive examples of "areas of investigation which are related to language structure as presented here, but remain to be undertaken." Others may be added, such as for example the

29. John Corcoran has been developing some interesting work on these lines, which he had intended to publish here. See Corcoran (1971a–c) for antecedents.

retention of reduced traces of sentence-intonation and lexical stress in a paratactic conjunct (interruption) and in the further reductions that result in various modifiers; the relation of sublanguages to one another and to less well-specifiable kinds of language use in a 'whole language'; problems of language variation and change, areal phenomena, and language typology, seen from this perspective; the structure of discourse prior to the linearization of word-dependencies in sentences, the integration of such structures in language-borne aspects of knowledge and learning, and their relation to the nonverbal universe of perception. Broader questions arise: What is language, really, that it is so central to human nature? By what interplication of biological and social inheritance is it acquired, and how may it be sharpened and extended in the ongoing evolution of society, if not of species, in which we may perhaps take an increasingly active role?

Other researchers and new students take up these questions. What we see here is only a beginning.

References

Anderson, Stephen R. 2001. "Commentary on 'Why Cultural Anthropology Needs the Psychiatrist': Why Linguistics Needs the Cognitive Scientist". *Psychiatry* 64.1: 11–13.

Bloch, Bernard. "Phonemic overlapping". *American Speech* 16: 278–284.

Bloch, Bernard. 1953. "Contrast". *Language* 29: 59–61.

Borel, Emile Felix Edouard Justin. 1928. *Leçons sur la theorie des fonctions.* 3e ed. Paris: Gauthier-Villars & Cie.

Chomsky, Noam. 1957. Review of *Fundamentals of Language*, by Roman Jakobson & Morris Halle (The Hague: Mouton, 1956). *IJAL* 23: 234–242.

Chomsky, Noam. 1964. *Current Issues in Linguistic Theory.* The Hague: Mouton.

Chomsky, Noam. 1975. *The Logical Structure of Linguistic Theory.* New York: Plenum.

Chomsky, Noam. 1977a. *Dialogues avec Mitsou Ronat.* Paris: Flammarion.

Chomsky, Noam. 1977b. *Language and Responsibility. Based on conversations with Mitsou Ronat.* Transl. by John Viertel [with collaboration and extensive revision by N. Chomsky]. New York: Pantheon Books.

Chomsky, Noam & Morris Halle. 1968. *The Sound Pattern of English.* New York: Harper & Row.

Corcoran, John. 1971a. "Discourse grammars and the structure of mathematical reasoning I: Mathematical reasoning and the stratification of language". *Journal of Structural Learning* 3.1: 55–74.

Corcoran, John. 1971b. "Discourse grammars and the structure of mathematical reasoning II: The nature of a correct theory of proof and its value". *Journal of Structural Learning* 3.2: 1–16.

Corcoran, John. 1971c. "Discourse grammars and the structure of mathematical reasoning III: Two theories of proof". *Journal of Structural Learning* 3.3:1–24.
Goodman, Nelson. 1951. *The Structure of Appearance*. Third edition. Dordrecht & Boston: Reidel.
Gross, Maurice. 1975. *Méthodes en syntaxe*. Paris: Hermann.
Halle, Morris. 1954. "The strategy of phonemics". *Word* 10:197–209.
Harary, Frank & Herbert H. Paper. 1957. "Toward a general calculus of phonemic distribution." *Language* 33:143–169.
Harris, Zellig S. 1933. "Acrophony and vowellessness in the creation of the alphabet". *Journal of American Oriental Society* 53.387. [Summary of 1932 Univ. of Pennsylvania M.A. thesis "Origin of the Alphabet".]
Harris, Zellig S. 1936. *A Grammar of the Phoenician Language*. (=*American Oriental Series*, 8.) New Haven, Conn.: American Oriental Society. [Ph.D. dissertation, Univ. of Pennsylvania, Philadelphia, 1934.]
Harris, Zellig S. 1940. Review of Louis H. Gray, *Foundations of Language* (New York: Macmillan, 1939). *Language* 16.3:216–231. [Page numbers cited from repr. in Harris 1970:695–705 under the title "Gray's *Foundations of Language*".]
Harris, Zellig S. 1941a. "Linguistic structure of Hebrew". *Journal of the American Oriental Society* 61:143–167. [Also published as Publications of the American Oriental Society; Offprint series, No. 14.]
Harris, Zellig S. 1941b. Review of N. S. Trubetzkoy, *Grundzüge der Phonologie* (Prague, 1939). *Language* 17:345–349. [Page numbers cited from repr. in Harris 1970:706–711.]
Harris, Zellig S. 1944a. "Yokuts structure and Newman's grammar". *IJAL* 10.4:196–211. [Repr. in Harris 1970:188–208.]
Harris, Zellig S. 1944b. "Simultaneous components in phonology". *Language* 20:181–205. [Repr. in Joos 1957:124–138 and in Harris 1970:3–31.]
Harris, Zellig S. 1946. "From morpheme to utterance". *Language* 22.3:161–183. [Repr. in Harris 1970:100–125, and in Harris 1981:45–70.]
Harris, Zellig S. 1951[1946]. *Methods in Structural Linguistics*. Chicago: Univ. of Chicago Press. [Completed and circulated in 1946, Preface signed "Philadelphia, January 1947". Repr. with the title *Structural Linguistics* as "Phoenix Books" P 52, 1960.]
Harris, Zellig S. 1952a. "Culture and style in extended discourse". *Selected Papers from the 29th International Congress of Americanists (New York, 1949)*, vol. III: *Indian Tribes of Aboriginal America* ed. by Sol Tax & Melville J. Herskovits, 210–215. New York: Cooper Square Publishers. [Page numbers cited from repr. in Harris 1970:373–379].
Harris, Zellig S. 1952b. "Discourse analysis". *Language* 28.1:1–30. [Page numbers cited from repr. in Harris 1970:313–348].
Harris, Zellig S. 1954. "Transfer grammar". *IJAL* 20.4:259–270. [Page numbers cited from repr. in Harris 1970:139–157].
Harris, Zellig S. 1955. "From phoneme to morpheme". *Language* 31.2:190–222. [Repr. in Harris 1970.32–67].
Harris, Zellig S. 1962. *String Analysis of Sentence Structure*. (= Papers on Formal Linguistics, 1.) The Hague: Mouton.
Harris, Zellig S. 1965. "Transformational theory". *Language* 41.3:363–401.

Harris, Zellig S. 1968. *Mathematical Structures of Language*. (=Interscience Tracts in Pure and Applied Mathematics, 21.) New York: Interscience Publishers, John Wiley & Sons.

Harris, Zellig S. 1970. *Papers in Structural and Transformational Linguistics*. Ed. by Henry Hiż. (=*Formal Linguistics Series*, 1.) Dordrecht: D. Reidel.

Harris, Zellig S. 1981. *Papers on Syntax*. Ed. by Henry Hiż. (=*Synthese Language Library*, 14.) Dordrecht: D. Reidel.

Harris, Zellig S. 1982. *A Grammar of English on Mathematical Principles*. New York: John Wiley & Sons.

Harris, Zellig S. 1990. "La genèse de l'analyse des transformations et de la métalangue". *Langages* No.99 (Sept. 1990) 9–19. [Ed. by Anne Daladier; translation into French by A. Daladier of most of the essay that introduces the present volume.]

Harris, Zellig S. 1991. *A Theory of Language and Information: A mathematical approach*. Oxford & New York: Clarendon Press.

Harris, Zellig S. 1997. *The Transformation of Capitalist Society*. Baltimore: Rowman & Littlefield.

Harris, Zellig S., Michael Gottfried, Thomas Ryckman, Paul Mattick, Jr., Anne Daladier, Tzvee N. Harris, & Suzanna Harris. 1989. *The Form of Information in Science: Analysis of an immunology sublanguage*. Preface by Hilary Putnam. (= *Boston Studies in the Philosophy of Science*, 104.) Dordrecht & Boston: Kluwer Academic Publishers.

Hymes, Dell & John Fought. 1981. *American Structuralism*. The Hague: Mouton.

Lowie, Robert H. 1939. *Hidatsa Texts Collected by Robert H. Lowie, with grammatical notes and phonograph transcriptions by Z. S. Harris and C. F. Voegelin*. (= *Prehistory Research Materials*, 1:6), 173–239. Indianapolis: Indiana Historical Society. (Repr., New York: AMS Press, 1975.)

Nevin, Bruce E. 1993. "A minimalist program for linguistics: The work of Zellig Harris on meaning and information". *Historiographia Linguistica* 20.2/3:355–398.

Nevin, Bruce E. 1995. "Harris the Revolutionary: Phonemic theory". *History of Linguistics 1993: Papers from the Sixth International Conference on the History of the Language Sciences (ICHoLS VI), Washington, D.C., 9–14 August 1993*, ed. by Kurt R. Jankowsky, 349–358. Amsterdam & Philadelphia: John Benjamins.

Prince, Alan & Paul Smolensky. 1993. *Optimality Theory*. (=*Technical Reports of the Rutgers University Center for Cognitive Science*, RuCCS TR-2.) New Brunswick, NJ: Rutgers Center for Cognitive Science.

Ryckman, Thomas A. 1986. *Grammar and Information: An investigation in linguistic metatheory*. Unpublished Ph.D. dissertation, Columbia University, New York.

Salkoff, Morris. 1999. *A French-English Grammar: A contrastive grammar on translational principles*. Amsterdam & Philadelphia: John Benjamins.

Acknowledgements

The initiator of this project was Harris's friend and one-time student, William M. Evan, without whose encouragement and helpful suggestions at certain trying junctures it might have foundered. I am also most grateful to the other members of my advisory board, Henry Hiż, Henry Hoenigswald, and Hilary Putnam, for their counsel and assistance in many matters.

Stephen Johnson has played an extremely important role as an advisor and reviewer. Of signal importance were the suggestions, sometimes critical, always helpful, of many people, including especially Naomi Sager, Michael Gottfried, Tom Ryckman, Paul Mattick, and Jim Munz. My repeated efforts to persuade Noam Chomsky to contribute proved fruitless, but the process helped to sharpen the argument in portions of the Foreword. Anne Daladier provided the English original of Harris's essay, which introduces this volume, at a time when a copy had not been found here; and for this, and indeed for her role in encouraging Harris to write it in the first place, we all owe her a great debt of gratitude.

But the greatest acknowledgement and thanks must go to the contributors to these two volumes, and to others not directly represented here, who have carried forward the various lines of research initiated by Harris. If Zellig Harris were still with us, he would doubtless express his appreciation to them, and to his colleagues, especially those in the University of Pennsylvania Department of Linguistics, which he founded in 1946. He would probably also express his appreciation to generations of his gifted students, and acknowledge how discussions with them helped him to develop and clarify his ideas.

Finally, he would undoubtedly express appreciation to his wife, Bruria Kaufmann, for her many substantive contributions over the years, and to his brother, Tsvee Harris and his wife Susannah, whose knowledge of the field of immunology was an invaluable contribution to the research that resulted in *The Form of Information in Science*.

Bruce Nevin

The background of transformational and metalanguage analysis

Zellig S. Harris
The University of Pennsylvania

This chapter provides a record of the background and the steps of analysis that led to grammatical transformations and to the recognition of the metalanguage as being a part of natural language, with the ensuing development to an operator-argument theory of language.

The work began as an attempt to organize the analyses made in descriptive linguistics, and to specify and formulate its methods. The later steps were called into being by this work of specification. The background for the work came largely from the foundations of mathematics and logic, and the analysis of formalisms; this was relevant to language because in all of these systems there were sentences (propositions, formulas) with partially similar structure (syntax). More specifically, there was the then current constructivism: in the criticisms by the intuitionists in mathematics (L. E. J. Brouwer), in Russell's theory of types, in the work of Emil Post, and in the Turing Machine procedure. Later, I considered that I had philosophical support in the constructivist (nominalist) approach of Nelson Goodman's *The Structure of Appearance*. Going back to the early background, there was also the development of recursive methods by Gödel, Tarski, and others; and in logic there was J. Łukasiewicz' Sentential Calculus, S. Leśniewski's Categorial Grammar, and the syntax of logic in W. V. O. Quine's *Mathematical Logic* of 1940.

Within linguistics, the success of de Saussure's phonemic analysis showed the usefulness of complementary and free variation as a basis for defining more unrestricted entities, stated as having various alternate values that were usually more restricted. In morphology and syntax, the 'distributional' method followed by Franz Boas, and more explicitly by Edward Sapir and Leonard Bloomfield, analyzed likewise the occurrence and combination of grammatical elements in the particular environments of other elements. I think, and I am

glad to think, that the intellectual and personal influence of Sapir and of Bloomfield colors the whole of the work that is surveyed below. It seemed natural to formulate all the methods above in the spirit of the syntax of mathematics and logic noted here.

This methodological program involved finding the maximum regularity in the occurrence of parts of utterances in respect to other parts. In its most general form it required the description of the departures from randomness in the combinations of elements, i.e. the constraints on freedom of occurrence of elements in respect to each other. Although decades of work were needed for applying the methods, and for further directions that grew out of the book, it is indicative of the intellectual background cited above that the general program could be stated from the beginning, e.g. in a paper in the *Journal of the American Oriental Society* 61 (1941) pp. 143, 166; also in "The Phonemes of Moroccan Arabic," *ibid.* 62 (1942) Sec. 4; (*Methods in*) *Structural Linguistics*, p. 364 (the latter was completed and circulated in 1946, though it appeared only in 1951). Some of the papers cited here are also in *Papers in Structural and Transformational Linguistics*, Reidel 1970; also *Papers on Syntax*, Reidel 1981.

As originally applied by the distributional linguists, the method involved collecting complementarily-, or similarly-, combining entities into a class, and then defining the class as a new 'higher' entity. The higher entity is in general more unrestricted (has greater freedom of occurrence) than the entities which it classifies. In addition, sequences of entities, or of these higher entities, may be found to constitute more regularly or freely occurring entities. All these classifications may be repeated to compose a hierarchy of higher entities (*Structural Linguistics* p. 369).

Formulating this hierarchy of distributional classifications, which came to be called structural linguistics, made it necessary to establish procedures for determining the primitive elements at the bottom of the hierarchy, for their simplicity and objective characterizability is as important to the system as are the classifications and sequences that state the departures from randomness of the entities at each level. Thus stated, the final system is finitary, in the sense of S. C. Kleene, *Introduction to Metamathematics* (1952).

In particular, the pair test of sound discrimination among speakers of a given language was offered as a basis for phonemic distinctions (*Structural Linguistics* p. 32), these being the ultimate and necessarily discrete primitives of the language structure. Phonemes were defined as a convenient arrangement of how these phonemic distinctions appeared in utterances, in complementary or free alternation. Morphemes and word boundaries were then

obtained by a stochastic process on phoneme sequences in utterances; they were the points of least restriction in the phoneme sequence ("From Phoneme to Morpheme," *Language* 31 (1955) 190–222; the method had been presented at Indiana University in the 1940s, and its possibility is recognized in *Structural Linguistics* p. 363).

In going on to morphology and syntax, there was a conscious effort to apply, mutatis mutandis, the methods that had been famously successful for the phonemic substrate. This program was carried out in "Morpheme Alternants in Linguistic Analysis," *Language* 18 (1942) 169–180, and a number of following papers. Thereafter no a priori justifiable general method was found to reach the structure of a sentence (or an utterance) by a hierarchy of constituent word sequences, or other partial structures of words. The problem was finally resolved by a single general procedure of building, around certain words of a given sentence, graded expansions in such a way that the sentence was shown to be an expansion of a particular word sequence in it, this word sequence being itself a sentence. (The first application of the method was in Hidatsa, a Siouan language; then an application to a Dravidian language was published in "Emeneau's Kota texts," *Language* 21 (1945) 283–289; the full method was presented at the Linguistic Institute, and appeared together with the Hidatsa analysis in "From Morpheme to Utterance," *Language* 22 (1946) 161–183.

The relevance of the hierarchy of word expansions, which was organized into an ascending chain of equivalences, was not simply in providing a direct procedure that yielded the structure of a sentence in terms of its words, but in opening a general method for the decomposition of sentences into elementary sentences, and thus for a transformational decomposition system. This unexpected result comes about because, first, the small sentence which is at the base of the expansions is recognizable as the grammatical core sentence of the given sentence, and, second, each expansion around a particular word can be seen to be a reduction or deformation of a component sentence within the given one. The status of expansions as component sentences was visible from the beginning: when the expansion method was presented at the Linguistic Institute, a question was raised as to how the method would distinguish the two meanings of *She made him N* in *She made him a good husband because she made him a good wife*; the answer was in showing that two different expansions obtained from two different component sentences yielded here the same word sequence (sec. 7.9 in the paper cited above).

The expansion analysis was formulated later as a decomposition of the given sentence into word strings ("Computable Syntactic Analysis," *Transfor-*

mations and Discourse Analysis Papers 15 (1959); *String Analysis of Sentence Structure* (1962)). The string status of the words of a sentence then made possible a stochastic procedure for finding, in the word sequence of an utterance, points of least restriction which were the sentence boundaries in that utterance (*Mathematical Structures of Language*, Interscience Tracts of Pure and Applied Mathematics 21 (1968) pp. 36–40).

While the machinery for transformations was provided by the "Morpheme to Utterance" equivalences, the motivation for developing transformations as a separate grammatical system was furthered by the paraphrastic variation in sentences that was found in discourses. In 1946, with the completion of *Methods in structural linguistics*, the structure of a sentence as restrictions on the combination of its component parts seemed to have gone as far as it could, with the sentence boundaries within an utterance being the bounds of almost all restrictions on word combination. I then tried to see if one could find restrictions of some different kind which would operate between the sentences of an utterance, constraining something in one sentence on the basis of something in another. It was found that while the grammatical structure of any one sentence in a discourse was in general independent of its neighbors, the word choices were not.

In a discourse, the component sentences revealed by the Morpheme to Utterance expansions were often the same sentence appearing in different paraphrastic forms in neighboring sentences. The use of reductions and deformations of sentences both to produce expansions and also to produce separate paraphrastic forms of a sentence motivated the formulation of a whole transformational system; and a list of English 'grammatical transformations' was included in the report presented to the Linguistic Society of America in 1950 ("Discourse Analysis," *Language* 28 (1952), 1–30, sec. 2.33). The transformational system (sketched below) was presented to the Linguistic Institute at Indiana University in 1951–1952. The formal presentation, with detailed structural-linguistic evidence that the expansions were indeed transformed sentences, was given at the Linguistic Society meeting in 1955 ("Co-occurrence and Transformation in Linguistic Structure," *Language* 33, (1957) 289–340).

In those years I had conversations about transformations with many people: with Piaget, and the psychologist David Rapaport, with Carnap and his follower Y. Bar-Hillel, with Max Zorn (of the lemma) to whom I showed the whole system at the Indiana Linguistic Institute, and with others. I had many very helpful discussions with Henry Hoenigswald, M. P. Schützenberger, and Maurice Gross, and enlightening comments from André Lentin. My closest

work was with Henry Hiż, who did a great amount of work on the methods, especially such as brought in considerations from mathematical logic, e.g. in "Congrammaticality, Batteries of Transformations, and Grammatical Categories," in *Proceedings of the Symposium of Applied Mathematics*, American Mathematical Society, 12 (1961) 43–50; "The Role of Paraphrase in Grammar," *Monograph Series on Language and Linguistics*, Georgetown University 17 (1964); also on establishing zeroings and transformations by eliciting sentence-completions and the like from speakers of the language. I also had a great many conversations with my students, above all with Noam Chomsky, who moved on in the direction of a comprehensive generative transformational system in his *Syntactic Structures* (1957), "A Transformational Approach to Syntax" in A. A. Hill, *Proceedings of the Third Texas Conference on Problems of Linguistic Analysis in English 1958* (U. of Texas 1962), and many major later books.

The consideration of paraphrase in discourse occasioned considerable investigation as to what should be stated as the criterion for one's saying that a difference between two sets of sentences is to be counted as a transformation from one set to the other. In the early years, after the initial success of the passive as a transformation, I had tried unsuccessfully such pairs as *I sold books to him* / *He bought a book from me* and *I lost a game to him* / *He won a game from me* (but *I lost the book* / *He found the book*); and somewhat differently *He cut through the wrapping with a knife* / (*In my hands,*) *the knife cut through the wrapping*. The solution, which was tested (with many people) on many sentence sets was the preservation of the grading of speakers' acceptance for the same word choices in two sets of sentences, preferably otherwise recognized sets (e.g. *The cat drank the milk* / *The milk was drunk by the cat* as vs. *The cat drank the word* / *The word was drunk by the cat*). The non-transformational status of most synonyms (e.g. *oculist* / *eye doctor*) was established then (whence the discussion in "Distributional Structure," *Word* 10 (1954) 146–162, sec. 2.3).

An important factor in thinking of a transformational system that would isolate structural paraphrase was the Skolem normal form in logic, which made me think of the possibility of a canonical form for sets of paraphrastic sentences. Transformations thus came out as paraphrastic equivalence relations among sentences. They were called transformations (rather than deformations, or other terms I had considered) because they were partial transformations in the set of sentences, mapping sentences in one subset onto same-word-choice sentences in the other, thus preserving word choice. The transformations provide a decomposition in the set of sentences, and the ultimate elementary sentences were called sentences of the kernel, because given the factor set (the set of

sentences over the set of transformations), then in the natural mapping of the set of sentences onto its factor set, the sentences which are sent into the identity of the set of transformations are the elementary sentences of the language.

Thus defined, the transformational system had various convenient features. It had a constructivist character, since it was possible to state the applicability of a transformation in terms of the last event in the construction of the sentence on which the transformation acted, so that transformations become steps in the construction of a sentence, rather than only phenomenological relations between sets of sentences. It was also found that many relations and mappings of sentence sets could be defined on the basis of transformations. While these were of little interest as applied mathematics, they were relevant in that almost every class and event in a given sentence structure could be defined by such relations, operations, and mappings, beginning from the elementary sentences which were the components of the given one.

Before leaving the subject of transformations, we consider the issue of generating (or deriving, synthesizing, predicting) as against analyzing (or describing, recognizing the structure of a sentence), which issue is commonly associated with transformations. First, the course of construction of a sentence suggests how a sentence can be considered to be derived from (expanded from) component sentences, with corresponding contributions of meaning to the source sentences. The transformational history of a sentence suggests derivation more strongly, because the derived sentence remains largely paraphrastic, hence 'the same sentence' in meaning; though in fact the 'derived' sentence is not derived in some general sense from its source, but merely contains the source (with expansions or shape changes). The well-known generative transformational theory of Noam Chomsky produces constituent components ('phrase structure') of the sentence and also its (later) transformations; the tree representation there could be considered a representation not so much of the sources of the sentence as of the ordered choices to be made in that system for producing the given sentence.

However, the difference between the analysis of a sentence and its generation is not substantive for the theory except in a limited but important sense (below), but rather is a matter of presentation. The analyzing of a sentence in structural linguistics allows both for a description which directly recognizes the structure, and alternatively for a grammar as a deductive system that synthesizes (generates) sentences (*Structural Linguistics*, pp. 365, 372; "Transfer Grammar," *International Journal of American Linguistics* 20 (1954: 259–270), or for transformationally generating it as noted in the "Co-occurrence

and Transformation" paper (sec. 5.6) cited above. In the latter case, the analytic statements of successively entering components of a sentence, or its decomposition, can be used almost directly to generate or predict sentences of that structure. In any case, analysis of the language precedes synthesis.

The difference between analysis and generation is rather that analysis has to identify as far as possible every regularity in speech or writing, and above all to recognize degeneracies, whereas generating can be done with just enough information about the language to distinguish in a general way every utterance from those not systematically identical with it (*Structural Linguistic* pp. 365–366). Analysis thus faces considerable additional difficulties. This lesser burden in generating becomes relevant when the informational features of grammar are distinguished from the non-informational ones (as in operator-grammar theory). For then, if all we want is to supply the information in a discourse, one method might be to generate an informational representation of its sentences, using the informational features alone (except insofar as some non-informational features are grammatically required or desirable in certain situations). The operator-argument theory below is in certain respects a generative theory with this capability.

We now consider the point about the metalanguage being in the language. First, as one analyzes the sentences of a language in more systemic detail one finds various evidences that metalinguistic sentences exist and operate in the grammar of a language. For example, the constraints on word choice in conjoined sentences (SCS) are such that generally the SCS is less immediately acceptable if there is no word that occurs in both component S. If nevertheless a given SCS is acceptable without this word repetition, it is found that there exists some sentence which would complete the repetition (by containing some word from each of the component S), creating an SCSCS which satisfies the condition. What creates the given SCS is that here the added S is zeroable (by the established low-information criterion for zeroing), either as being known to the hearer or as being a dictionary definition, or the like, known to all users of the language (*Mathematical Structures of Language* pp. 131–138). This shows the existence of metalinguistic sentences, giving the grammar and dictionary of the sentence to which they are adjoined.

Second, the explicit structure of statements in logic and mathematics had made it clear that the statements about this structure could not be expressed within this structure: the metalanguage of mathematics was outside mathematics. (See for example Alonzo Church, *Introduction to Mathematical Logic*, Princeton U. Press, 1956. While the term 'metalanguage' as used in the lin-

guistic work is an extension of the use in Rudolf Carnap, *The Logical Syntax of Language*, it also satisfies the more stringent (finitary) condition for the term 'meta' in S. C. Kleene *Introduction to Metamathematics*.) The structure of the metalanguage had been left undescribed, the view being that it, or its metalanguage in turn in infinite regress, has to be undescribed and indeed not fully specifiable, simply given in natural language. This conforms to the common view in philosophy that natural language is amorphous, or in any case not fully specifiable.

However, as the analysis of natural language showed it to be specifiable in as great detail as we wish, with the unspecified residue being encapsulated as structurally secondary in respect to the main description, it became possible to specify the structure of the statements about natural language in comparison with the sentences that these sentences describe. In the first place, the metalinguistic statements are themselves sentences of natural language. In the second place, they are a structurally specifiable subset of those sentences and constitute a sublanguage in the sense given below ("Algebraic Operations in Language Structure," International Congress of Mathematicians, Moscow 1966; *Mathematical Structures of Language*, pp. 17, 125–128).

The motivation for studying the metalanguage of natural language was not only its different status from that of the metalanguage of mathematics, but also its value in formulating the syntax of natural language. Several important clarifications in syntax are achieved by specified relations between sentences of language and certain metalinguistic sentences adjoined to the sentences of which they speak. The most important is a single derivation for tense, which yields both "absolute" and "relative" tenses (i.e. tense in respect to time of speaking and tense in respect to neighboring sentences) and also non-time uses of tense (*Notes de Cours de Syntaxe*, Maurice Gross, tr. and ed., Editions du Seuil, 1976: 158–181). Another is a derivation of reference from cross-reference, in a way that explains, for example, why the "free" pronouns (e.g. *he*) have no fixed location for their antecedent (*A Grammar of English on Mathematical Principles*, Wiley 1982: 87–97). More generally, the metalanguage makes the whole of a natural language self-contained; and each sentence in the language becomes self-contained when we adjoin to it the zeroable metalinguistic sentences which state the meaning and grammatical relations of each morpheme in it.

A crucial methodological contribution of the metalanguage is the following: since it is impossible to define the elementary entities and constraints of a language by recourse to its metalanguage (since the metalanguage is itself

constructed from those entities by means of those constraints), it follows that the structure of language can be found only from the non-equiprobability of combinations of parts. This means that the description of a language is the description of contributory departures from equiprobability, and the least statement of such contributions (constraints) that is adequate to describe the sentences and discourses of the language is the most revealing.

Whereas the issues considered before this point were methods of analysis for language, the operator-argument construction suggests not only a method but also a theory. This may therefore be the place for a brief excursus about a theory in a field such as linguistics. Differently from most sciences, linguistics admits of an alternative to theory: an orderly catalog of the relevant data, sufficient to do most of the work that a theory is supposed to do. Both the knowledge of what is relevant in data and the possibility for orderly coverage are due to the recognition of a finite set of phonemic distinctions (hence, discrete phonemes) in each language, and also to the fact that only statable phoneme sequences (finite in number and length) constitute morphemes or words — the material from which syntax is directly made. Hence, the relevant information about a language can be given most completely and interestingly by a listing of each morpheme, each with a complete listing of the properties (its combinabilities with other morphemes, and its forms in those combinations): cf. the work being done by Maurice Gross and his associates (note *Structural Linguistics* pp. 376). Sophistication as to the properties that should be stated (grammatical class, selection, morphophonemics, transformations) may come from global investigation of regularities. But the coverage is assured by the listing, though one may fail to restore such properties as morpheme presence in zero shape, and though some properties may have such large and regular morphemic domain as to make wasteful a listing by individual morpheme.

Furthermore, given what we know about the status of 'truth' in logic and about the alternative description of structure, a theory should not be thought of as presenting the final truth, but only as organizing the results of certain methods of analysis, 'true' as far as it goes.

Nevertheless, theoretical formulations present global results that are not (or not directly) represented in a catalogue of relevant data:

– The fuzziness of some domains.
– The essential relation among properties as it comes out in a deductive system.

- The relevant distinction of partially-ordered and linearly-ordered processes in language.
- The relating of forms and combinations to processes (some recursive) which create those forms and combinations.
- The relation of structure to change and then to the development of language.
- The basic relation of language structure, and language development, to expressing and transmitting information.

In the case of language, there is also a responsibility to formulate a theory based on self-organizing capacities: one that will present language as a system that can arise in a state in which language does not exist. This is so because of the unique status of language as a system which contains its own metalanguage. Any description we make of a language can only be stated in a language, if even only in order to state that some items of the description are properties of some other items (i.e. how to read the table). We cannot describe a language without knowing how our description can in turn be described. And the only way to avoid an infinite regress is to know a self-organizing description, which necessarily holds also for the language we are describing even if we do not use this fact in our description.

In addition to its service in analyzing sentences, transformational methods as used here had shown another way of characterizing a sentence: by the subset of sentences to which it belongs. This arose from the various sentence relations, more exactly sentence subset relations, and the mappings and partitions, in transformational analysis. The next question was naturally what other kinds of relevant and useful subsets of sentences might be found. The detailed study of the structure of language then showed that one can define within a natural language various sublanguages, such as the metalanguage, and the language use in many sciences (*Mathematical Structures of Language*, pp. 152–155). These were subsets of the sentences of the language which had the properties of sublanguages. It was not really a matter of subsets of the vocabulary: any subset of sentences or of discourses in a language would contain only a small part of the vocabulary of the language. What is special to sciences is that certain subclasses of words (and phrases) co-occur in a regular way to make certain specifiable sentence-types (as sentential combinations of word subclasses); and here, as the corpus of material being investigated grows, the set of sentence-types grows little if at all. If we take any two sentences of the subset and operate on them with various operators of the whole language, e.g. *and* or various

transformations, we obtain again a sentence of the subset. Hence the subset is a sublanguage. The grammar of the sublanguage, however, can be shown to be not a subset of the grammar of the whole language; rather, it intersects, importantly, the latter grammar. The study of sublanguages was not influenced by any particular situation outside linguistics, except for the general example of the existence of subsystems in mathematical systems. An example of a sublanguage is given in Harris, Z., Gottfried, M., Ryckman, T., Mattick, P., Daladier, A., Harris, T. N., Harris, S., *The Form of Information in Science: Analysis of an Immunology Sublanguage* (Kluwer Acad. Publ., Dordrecht, 1989).

There was finally a motivation that led out of transformations to an operator-argument theory of language. This came from two directions. First, for all that they contributed to linguistic theory, transformations were not general enough (their conditions were too specific, and the sequential applicability in deriving sentences was too limited), and not elementary enough (there were too many to constitute a reasonable set of primitives for a new 'derivational' dimension of sentence construction). Second, some of the more complex transformations (e.g. the English passive) were physically (morphemically) identical to successive application of particular added morphemes; and their domain restriction equalled the selection restrictions of just those morphemes ("The Elementary Transformations," *Transformations and Discourse Analysis Papers* 54 (1964); "Transformational Theory," *Language* 41 (1965) pp. 363–401). This led to defining a system of reductions in sentences such that most transformations are either a reduction or a successive application of reductions. It was then found that in most sentences the ultimate source sentence which was the starting point for the reductions had a simple subject-predicate or subject-verb-object structure (i.e. an operator-argument structure). The importance is not in this observation by itself, which is close to the popular understanding of language (and to logic from Aristotle and on), but in the ability to obtain other sentences from these simple predications by usually a priori statable reductions and recursively stated predications upon predications. Both the reductions and the predications upon predications were constrained to occur only in stated conditions. In the case of reductions, it was possible to say that they occurred when one participant word in an operator-argument structure had exceptional likelihood of occurring there in respect to the other participant word (hence the reductions were paraphrastic, contributing little or no information to the sentence at that point).

In line with the general 'distributional' program, this analysis was first used for sentences that involved restrictions, where the operator-argument

source was found to be less restricted (because the domains of successive reductions are monotonic descending since each reduction can only be stated on all or a proper part of the set of words in the given operator-argument position). However, it was found that most of the remaining non-operator-argument sentences could be derived from operator-argument sentences by the same reductions and predications-upon-predications that had been used for the restricted sentence structures. The specific program to arrive at an operator-argument source was first used for the English comparative, in a way that incidentally explained Sapir's point that, e.g., *He is richer* does not imply: *He is rich* ("Grading: A Study in Semantics," *Philosophy of Science* 11 (1944) pp. 93–116; *Selected Writings of Edward Sapir*, D. G. Mandelbaum, ed. U. of Calif. Press 1958: 122–149). The English comparative analysis is sketched in *Mathematical Structures of Language*, pp. 174–175; although this book was an expansion of a lecture given at the Courant Institute of Mathematical Sciences in 1961, the analysis of the comparative is later and was inserted shortly before publication of the book.

What came out finally was a system of predicates (operators) first on primitive arguments and then recursively on predicates, with reduction of words which had high likelihood in the given operator-argument relation. This created a partial order of words in each sentence, and in the language as a whole. It was constructive, not only in the partial order of entry of words into a sentence, but also in that the reductions took place in a word upon its entry, so that each sentence could be defined as a particular kind of semi-lattice of word-occurrences and reductions. All further events in forming a sentence are defined on resultants of the partial order construction.

This method went beyond transformations in two respects. First, transformations brought word-choice specification into grammar by accepting the unspecified word-choice in elementary sentences and then preserving it under transformation. Operator grammar uses likelihood information about word-choice both in the first-level operators that create the elementary sentences and in the similarly-working second-level operators that create from them the enlarged sentences. Second, the operator grammar gives a single system of word partial order for forming both elementary sentences and other sentences (some of the latter called transformational), with reduction in shape to form the remaining sentences.

An excursus in word-choice: traditional and structural linguistics found regularities of co-occurrence word classes, between one or another word of another set. Because almost every word was unique in its selection, i.e. in what

individual words it occurred with most frequently, pre-transformational linguistics did not deal with the co-occurrence properties of individual words. In transformational analysis, it turns out that the undescribed relation has to be stated largely in the "kernel" sentence alone, and is merely preserved in the transforms. In the operator-argument theory, the selection of individual words in respect to their operators or their arguments, in whatever detail determined, constitutes the basic data. Structure is then created by the stable distinction between zero and non-zero possibilities of co-occurrence in the operator-argument relation, which creates syntax, and by the somewhat less stable high probabilities of co-occurrence that characterize the meanings of individual words. Transformations are then found to be largely reductions of form in the highest-probability (lowest information) co-occurrences. And structure beyond single sentences (conjunctions, discourse, sublanguage) is made by various constraints in the preservation of word choice within operator-argument relations.

The operator grammar reveals a sharper relation between the structure of a sentence and its information (as had been sought by Carnap and the Vienna Positivists); this, by specifying and ordering each departure from equiprobability. Some of these departures (mostly universal ones) can be seen by their structure to be information bearing (in a sense related to that of mathematical information theory); others (mostly in particular languages or language families) create regularities and irregularities that are not substantively informational. By-products of the whole theory are the status of the operator-argument system as a mathematical object (a necessity for the stability of language structure), and the picture of language as a self-contained, self-organizing, and evolving system (*A Theory of Language and Information: A mathematical approach*, Oxford U. Press, 1991). All of these properties, including the lack of an external metalanguage, can be looked upon as being expectable, given the use (function) of language, and the fact that it was not created in any conscious plan.

References

Ajdukiewicz, Kazimierz. 1978. *The Scientific World-perspective and other Essays*, 1931–1963. Dordrecht: D. Reidel.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, Noam. 1962. "A Transformational Approach to Syntax" in: A.A. Hill, *Proceedings of the Third Texas Conference on Problems of Linguistic Analysis in English*, 1958. U. of Texas.

Church, Alonzo. 1956. *Introduction to Mathematical Logic*, Princeton, NJ: Princeton University Press.

Goodman, Nelson 1951. *The Structure of Appearance*. Third edition. Dordrecht & Boston: Reidel.

Harris, Zellig S. 1941. "Linguistic Structure of Hebrew". *Journal of the American Oriental Society* 61.143–167. [Also published as Publications of the American Oriental Society; Offprint series, No.14.]

Harris, Zellig S. 1942a. "Morpheme Alternants in Linguistic Analysis". *Language* 18.3:169–180. (Repr. in Harris 1970:78–90, and in Harris 1981.23–35.)

Harris, Zellig S. 1942b. "Phonologies of African Languages: The phonemes of Moroccan Arabic". *Journal of the American Oriental Society* 62.4:309–318. (Repr. as "The Phonemes of Moroccan Arabic" in Harris 1970:161–176.)

Harris, Zellig S. 1945. "Review of Murray B. Emeneau, Kota Texts, vol. I" (Berkeley: Univ. of California Press, 1944). *Language* 21.283–289. (Repr. as "Emeneau's Kota Texts" in Harris 1970:209–216.)

Harris, Zellig S. 1946. "From morpheme to utterance". *Language* 22.3:161–183. (Repr. in Harris 1970:100–125, and Harris 1981:45–70.)

Harris, Zellig S. 1951 [1946]. *Methods in Structural Linguistics*. Chicago: Univ. of Chicago Press. (Repr. 1960 as *Structural Linguistics*, Phoenix Books; 7th impression, 1966; 1984.) [Completed and circulated in 1946, Preface signed "Philadelphia, January 1947".]

Harris, Zellig S. 1952. "Discourse Analysis". *Language* 28.1:1–30.

Harris, Zellig S. 1954a. "Transfer Grammar". *IJAL* 20.4:259–270. (Repr. in Harris 1970:139–157.)

Harris, Zellig S. 1954b. "Distributional Structure". *Word* 10.2/3:146–162. (Also in *Linguistics Today: Published on the occasion of the Columbia University Bicentennial* ed. by Andre Martinet & Uriel Weinreich, 26–42. New York: Linguistic Circle of New York, 1954. Repr. in Harris 1970:775–794, and in Harris 1981:3–22.)

Harris, Zellig S. 1955. "From Phoneme to Morpheme". *Language* 31.2:190–222. (Repr. in Harris 1970:32–67.)

Harris, Zellig S. 1957. "Co-Occurrence and Transformation in Linguistic Structure." *Language* 33.3:283–340. (Repr. in Harris 1970:390–457, and Harris 1981:143–210.

Harris, Zellig S. 1959. "Computable Syntactic Analysis". *TDAP* 15. Philadelphia: Univ. of Pennsylvania. (Revised as Harris 1962; excerpted, with the added subtitle "The 1959 computer sentence-analyzer", in Harris 1970:253–277.)

Harris, Zellig S. 1962. *String Analysis of Sentence Structure*. (= Papers on Formal Linguistics, 1.) The Hague: Mouton. [Revision of Harris 1959.]

Harris, Zellig S. 1964. "The Elementary Transformations". *TDAP* 54. Philadelphia: Univ. of Pennsylvania. (Excerpted in Harris 1970:482–532, and abbreviated in Harris 1981:211–235.)

Harris, Zellig S. 1965. "Transformational Theory". *Language* 41.3:363–401. (Repr. in Harris 1970:533–577, and in Harris 1981.236–280.)

Harris, Zellig S. 1966a. "Algebraic Operations in Linguistic Structure". Paper read at the International Congress of Mathematicians, Moscow. (Published in Harris 1970:603–611.)

Harris, Zellig S. 1968. *Mathematical Structures of Language.* (=Interscience Tracts in Pure and Applied Mathematics, 21.) New York: Interscience Publishers, John Wiley & Sons).

Harris, Zellig S. 1970. *Papers in Structural and Transformational Linguistics.* Ed. by Henry Hiż. Dordrecht/Holland: D. Reidel.

Harris, Zellig S. 1976. "On a theory of language". *Journal of Philosophy* 73.253–276. (Excerpted in Harris 1981.377–391.)

Harris, Zellig S. 1981. *Papers on syntax.* Ed. by Henry Hiż. (=Synthese Language Library, 14.) Dordrecht/Holland: D. Reidel.

Harris, Zellig S. 1982. *A grammar of English on Mathematical Principles.* New York: John Wiley & Sons.

Harris, Zellig S. 1991. *A Theory of Language and Information: A mathematical approach.* Oxford & New York: Clarendon Press.

Harris, Zellig S., Michael Gottfried, Thomas Ryckman, Paul Mattick, Jr., Anne Daladier, Tzvee N. Harris & Suzanna Harris. 1989. *The Form of Information in Science: Analysis of an immunology sublanguage.* Preface by Hilary Putnam. (=Boston Studies in the Philosophy of, Science, 104.) Dordrecht/Holland & Boston: Kluwer Academic Publishers.

Hiż, Henry. 1961. "Congrammaticality, batteries of transformations, and grammatical categories", *Proceedings of the Symposium of Applied Mathematics*, American Mathematical Society, 12:43–50.

Hiż, Henry. 1964. "The role of paraphrase in grammar", *Monograph Series on Language and Linguistics*, Georgetown University, 17.

Kleene S.C. 1950. *Introduction to Metamathematics.* New Jersey: D. Van Nostrand Co.; Amsterdam: North-Holland Publishing Co.

Leśniewski, Stanislaw. 1929. "Grundzüge eines neuen Systems der Grundlagen der Mathematik." *Fundamenta Mathematicae* 14:1–81.

Łukasiewicz, J. 1963[1929]. *Elements of Mathematical Logic*, Oxford: Pergamon Press; & New York: The MacMillan Company. Tr. by Olgierd A. Wojtasiewicz of *Elementy logiki matematycznej.* 1929. PWN, Warszawa.

Łukasiewicz, J. & A. Tarski. 1956[1930]. "Investigations into the sentential calculus". In Tarski 1956[1930]:38–59. Tr. by J.H. Woodger of "Untersuchungen über den Aussagenkalkül", *Comptes rendus de la Société des Sciences et des Lettres de Varsovie*, Classe III, 23(1930):30–50.

Quine, Willard van Orman. 1940. *Mathematical Logic.* Cambridge, Massachusetts: Harvard University Press.

Sapir, Edward. 1944. "Grading: a study in semantics", *Philosophy of Science*, 11 (1944) p. 93–116. Repr. in *Selected writings of Edward Sapir*, D.G. Mandelbaum ed., U. of Calif. Press, 1949, pp. 122–149.

Tarski, Alfred (ed.). 1956. *Logic, Semantics and Metamathematics.* Oxford & New York: Clarendon Press.

Part 1

Philosophy of science

Chapter 1

Method and theory in Harris's Grammar of Information*

T. A. Ryckman
Wolfson College, Oxford

1. Introduction

For more than a quarter century, the recent history of linguistics has featured a dramatic narrative relating how, beginning in the early 1960s, structuralism, as a viable research program, was rapidly eclipsed by the 'revolution' — in the classically Kuhnian sense of paradigm shift — of generative grammar. It appears irrelevant that American structuralism was never a paradigm in Kuhn's sense — a unified or even consistent program of methods, problems, goals or approaches — as any careful reading of the works of its two founders, Leonard Bloomfield and Edward Sapir, will quickly attest.[1] But the emphatic portrayal of a 'scientific revolution' has an obvious legitimating function in a discipline historically occupying nebulous and disputed turf abutting on the humanities, the natural sciences, and the social sciences. The narrative prominently features the engaging theme of Oedipal revolt. A small, but zealous, band of younger linguists, armed with the new tools of formal language theory, hypothetico-deductive methodology, and an incipient cognitive psychology, toppled a linguistic establishment wedded to methodologically crude inductivist and behavioristic 'discovery

*This chapter is dedicated to the memory of Zellig Harris, and is based on a paper given at a meeting of the Boston Colloquium for Philosophy of Science on October 8, 1991. My thanks to Michael Gottfried for commenting on that occasion, and to the members of the audience for the vigorous discussion that followed. Special thanks go to Bruce Nevin for all his assistance in reviving this chapter for the present volume.

1. On the different traditions stemming from Bloomfield and Sapir, see Hymes & Fought (1975).

procedures'. The resulting marriage of cognitive psychology and formal language theory, so the story goes, made possible for the first time a 'genuinely scientific' study of the nature of language and mind. According to the new orthodoxy, syntax and semantics in theoretical linguistics, no less than in the theoretical natural sciences, can proceed only by the method of 'conjecture and refutation', i.e., by fabricating 'bold hypotheses' *à la* Popperian anti-inductivism, to be tried upon 'sample data'. To anyone unschooled in the diversity of trends and directions within American structuralism, or indeed within any theoretical science, the fashionable metascience, reflecting contemporary trends in philosophy of science, appeared compelling. Within a few years the new linguistics, with its claim to be a modern revival of seventeenth century doctrines of innate ideas, attracted the critical scrutiny of the philosophers. The wider intellectual community soon followed, with the result that the new linguistics came to occupy, as the older had not, a prominent position on the stage of contemporary intellectual culture, debated even in the pages of *The New York Review of Books.* Over the years, however, much of the outside interest in the new linguistics waned, dismayed, certainly, by a confusing proliferation of alternative perspectives and factions, and ultimately, perhaps, by nagging awareness of lack of progress in narrowing the gap between striking claim and substantive result. Yet even in down-sized form, the new conception of linguistics, having severed its ties to formal language theory, and no longer self-described as 'generative', has largely retained its dominant image as the 'scientific linguistics', if only out of lack of awareness of a preferable alternative. In this respect, the legitimating function of disciplinary history has been confirmed: there has been no (or very little) looking back.

The linguistic theory of Zellig Harris is just such a neglected alternative. An outgrowth of structural linguistics, or rather the combinatorial (distributional) method he pioneered, Harris's views culminated in the last ten or so years of his life in a comprehensively synthetic account of the nature of language and information that is both remarkable in its detail and beautiful in the simplicity and power of its conceptual structure.[2] In its barest essence, Harris's work has made possible a detailed constructive mathematical consid-

2. See papers 12–16 (on Operator Grammar) in Harris (1981), Harris (1982, 1988), Harris et al. (1989), and the synthetic presentation, Harris (1991).

eration of language as a self-organizing system that expresses and carries information.[3]

2. Motivation

I'll begin by examining what I understand to be the guiding motivation for Harris's approach in linguistic theory. I do this in part because this motivation has been frequently misunderstood, hence overlooked, in part because of the intrinsic interest for philosophy of science of the intimate interconnection between his theory and methodology. The entry point for this discussion is the methodological postulate that there is no standpoint outside the data of language from which to advance theoretical inquiry. This is not to say that the study of language cannot or should not be enriched by results obtained across many and varied dimensions: social, psychological, historical, anthropological, acoustic, physiological, formal, semantic, and political, among others. Rather, the fundamental issue concerns the problem of investigating the structure of natural language, given that there is no metalanguage 'external' to natural language in which to define the elements of language and to characterize its grammar. Of course, we know pre-theoretically that such structures exist since not all combinations of linguistic elements are possible sentences or utterances in a given language; moreover, people can use and understand 'new' sentences and discourses of which they have had no previous experience. But in prohibiting appeal to an 'external' metalanguage, the intent is to forbid explanatory accounts of language structure that appeal to antecedent quasi-linguistic theoretical structures — logical, mathematical, semantic, conceptual — taken

3. The mathematical treatment of language arises from a fundamental relation of 'dependence on dependence' which partitions the set of words (rather, word occurrences in utterances) into dependence classes (argument-requirement classes) of operators and their arguments. Operator words have a direct analogy to sentence-forming functors in the categorial grammars of Leśniewski and Ajdukiewicz in that the satisfaction of any operator word's argument requirement yields a sentence (thus some words are only arguments, i.e., have null argument requirement). Two other primary constraints are defined in terms of the dependence on dependence requirement: 1) gross inequalities of likelihood obtaining between an operator word and the words of its argument class, and 2) reduction in the phonemic shape (perhaps to zero phonemic shape) of words which have a very high likelihood of occurring in a particular operator-argument environment (are 'expectable' in that environment).

as primitive and not, in contrast, themselves derivative from natural language and its structure. One does not have to be too well-versed in recent linguistics, philosophy of language, or philosophy of mind to find theories of language or meaning which violate this prohibition.[4] Obviously, for theories of this kind, there is no interesting notion of structure in language to be explored without the 'external' perspective. But should such a theory be adopted, the job of accounting for language structure becomes easier only if there is no obligation to explain the occurrence of the structures at the meta-level, in which case one is off and running into a regress. For Harris the threat of regress levied an obvious injunction against pseudo-explanation and a demand for an appropriate methodology that nowhere — even implicitly — requires an external standpoint. Although dim appreciation of this peculiar methodological situation facing the linguist was something of an article of faith in previous structuralism, I believe that Harris alone placed the proscription of an external standpoint at the center of the development of a proper linguistic methodology.[5]

The methodological situation facing the grammarian is precisely analogous to that of the logicist logician recognized earlier in the century by Bertrand Russell, and following him H. M. Sheffer who coined the inspired phrase, 'the logocentric predicament'. Fully realizing the ramifications of this 'predicament', Harris pioneered the development of the distributional, or, since this term led to misunderstandings and further misperceptions, 'combinatorial' (or combinational) approach to formulating the structure of language. This is just the task of formulating the constraints upon those combinations of linguistic elements that can occur, as attested by native users of the language or sublanguage. What Harris recognized before anyone else is that this method of characterization of language structure, in terms of successive levels of redundancy of combinations of elements, demonstrates — given the conceptual connection between redundancy and information[6] — that the cumulative effect of these intrinsic constraints is to create information. My present point,

4. Through appeal to levels of 'logical form', intensional logics or possible-world semantics, 'language of thought' hypotheses, *a priori* concepts and conceptual entailments, and so on.

5. Anecdotally: Once in conversation, when I had referred to him as a 'linguist', Harris demurred, disclaiming any title as linguist, and said he thought of himself as a 'methodologist'.

6. For example, a random set (where nothing is predictable), or a perfect crystal (where everything is predictable), intrinsically carry no information. On the conceptual connection between redundancy and structure, see Simon (1984).

however, is merely to call attention to Harris's assessment of the peculiar situation facing linguistic theory in view of the perceived nature of the object it studies, and, second, to underscore his conclusion that it is necessary to devise appropriate methods that nowhere depend upon an external standpoint.

For the moment the concern is 'logocentricity' and the predicament thereupon facing the linguist. What is the force of this perception? By way of illustration, we can return to the context in which Sheffer's remark was made, which concerned rather the plight of the logician: In a review of volume I of the second edition of *Principia Mathematica* in 1926, Sheffer wrote:

> [T]he attempt to formulate the foundations of logic is rendered arduous by a [. . .] 'logocentric' predicament. *In order to give an account of logic, we must presuppose and employ logic* (Sheffer 1926, emphasis in the original).

Sheffer is here expressing a view of logic as constitutive of any rational thought whatsoever. On pain of irrationality there could be no vantage point outside of logic from which to isolate and elucidate, let alone to critique, the foundational concepts of logic. This *sine qua non* conception of logic is that of Frege, found in somewhat different fashion in the early Russell and the Wittgenstein of the *Tractatus* (Ricketts 1979). Of course, it is no longer the modern one. Within just a few years after the *Tractatus* had cryptically proclaimed that 'illogical thought' was a kind of *contradictio in adjecto*, that logical form could only be 'shown', and that what can only be 'shown' cannot be 'said',[7] this logical monomania was completely undermined by the limitative results of Gödel and Tarski. Here distinctions are required between truth and proof, between expressibility and consistency in an object language and in its metalanguage. Intuitionism introduced a further axis of relativization in proposing new interpretations for even the logical constants. However, none of these 'relativizing' innovations impugned the force of the critique that the absolutist conception of logic directed upon its central target, psychologism. According to that complex tissue of doctrine, the concepts and laws of logic were rooted in psychological operations of the human mind that psychological investigation — in those days, that usually meant introspection — might eventually discover. Thus the logicist perception of a 'logocentric predicament' had an anti-psychological and anti-mentalistic mandate. Logic, the foundation of mathematics, is not reducible to psychology. Thanks to Frege's conception of a completely

7. Wittgenstein (1922), 5.4731 and 4.1212; cf. 6.123: "Clearly the laws of logic cannot in their turn be subject to laws of logic."

formal ('gapless') proof procedure for mathematics which transformed logic (at least after Russell and Whitehead's work), psychologicism in logic never recovered and logic became fully symbolic and 'mathematical'.

How does this episode bear on linguistics? There is at least an initial disanalogy which points to a crucial methodological difference between logic and grammar: the former but not the latter can presuppose the resources of language (as Frege, Russell, and Wittgenstein each observed). So there remains a fundamental distinction between the 'art of reasoning' and the 'art of speaking'.[8] Still, one might well conclude that perceptions of 'logocentricity' are simply the concomitant of militant anti-psychologism. In the case of structural linguistics, one can certainly read (parts of) Leonard Bloomfield in this way (whose anti-mentalism is coupled with, at times, an extreme behaviorism). Yet quite apart from the philosophical battle the logicists waged against psychologism, for American structural linguistics there was another, and quite different, avenue which led to a similar assessment of being placed in a 'logocentric predicament'. This arose within the Boas/Sapir tradition of anthropological linguistics and, in particular, from the study of Amerindian languages. Here, the concern was, as Boas warned repeatedly, not to prejudge or anticipate the description of these widely differing languages by forcing their description into a Procrustean bed of grammatical categories or paradigms inherited for the most part from the ancient Greek and Latin grammarians. Historians of ideas have noted a legacy of W. von Humboldt's conception of the 'genius' or character of languages in the 'particularistic' approach to the study of language by linguists of the Boas school, a conception, in the hands of B. L. Whorf, that issues in the strong thesis of linguistic relativity. My concern here is not with the claim of linguistic relativity (which certainly overreaches any evidence in its favor — Cf. Harris 1991: 387) but rather to call attention to the central figure of Sapir, both within this tradition and as influence on Harris, and to the force of his judgment that "Language is primarily a cultural or social product and must be understood as such" (Sapir 1929: 166).

---

8. In comments during the ensuing discussion of this chapter in 1991, Burton Dreben pointed out another disanalogy: for the logicists the very concept of an external standpoint was impossible — they lacked even the notion that logic had a grammar (here the contrast between the early and the later Wittgenstein is dramatic). Dreben also reminded us that another analogy to Harris's prohibition of an external standpoint is found in the central tendency of modern differential geometry, stemming from Gauss, of investigating intrinsic properties of figures and surfaces without reference to an embedding space.

In 1951, in a 45-page review of Sapir's papers (Harris 1951b: 288–333), Harris pointed out that, for Sapir, the major fact about language was the existence of patterns. The most important of these is sound pattern, but the salient point was that this patterning might best be understood as conventionally fixed aspects of cultural behavior in which the individual participated but which were, in an important respect, autonomous of the individual. (Sapir's writings on language 'drift' are a well-known example of these views.) To be sure, Sapir saw patterning in all aspects of culture and of individual participation in culture. But as Harris noted, the fundamental fact here was the discovery of the existence of such patterning; it was left to a later generation to establish more precisely the kinds of structure present and their relations to one another. The starting point, established by Sapir (and less clearly by Saussure) was the principle of the phoneme, the fundamental significance of which lies in a distinction between imitation and repetition. As sound spectography showed in the early post-war years, it is unlikely that any two utterances of the same sound are ever physically identical, yet each language has a relatively small number of functionally distinct sounds (perhaps two dozen or a few more) from which all utterances of the language are formed. A phoneme, of course, is an equivalence class of physical sounds, whose members perhaps differ widely in physical contours, yet any two members of which are accounted 'the same' sound among speakers of a language; i.e., any occurrence of one member is functionally a repetition of any of the others. The methodological import of Sapir's social and cultural conception of language can be located precisely here, in his emphasis on the expressly social character of the determination of patterning among sounds. This point is of surpassing importance for Harris's own innovative development of methods for investigating word combinations. For these can be seen as successively formulating broader and broader equivalence classes, elements of which, by virtue of their common environments of occurrence, are regarded as 'saying the same'. In effect, the method of phonemic analysis has been generalized and extended up through studies of languages in restricted semantic domains (e.g., science sublanguages) where word classes and subclasses, sentence types and even sentence sequence types are the equivalence classes of interest (see Harris 1989, passim). Harris was completely unique in combining the formal levels approach of Bloomfield with Sapir's intuitive and perceptive understanding of the inexorable linkage of language and culture, and moreover, in actually seeing no conflict between the explicit item-and-arrangement method of Bloomfield and Sapir's seemingly teleological 'process' formulations. The

larger point, however, is the emergence of a method for investigation of language structure which sees this structure as a social and cultural product both conditioning and conditioned by the aggregate of language users at a given time. As I will urge in a moment, this is a conception at considerable variance from the traditional view held in philosophy of language (from Descartes to Frege and beyond) which regards language primarily as a means for the expression of thought. And it indicates another route from the 'logocentric predicament' which stands independent of any particular anti-mentalist platform, such as Bloomfield's or Quine's, whatever the merits or demerits (and there are both) of such a critique. It is in this context then that we should view Harris's proscription of an external metalanguage.

3. Information

Thus far the motivation. We have now to take account of Harris's insight that pattern itself is accountable as a hierarchical structure of constraints upon combinations of elements, whose combined effect is accretional, and that, as a result, language structure 'carries', or better, expresses, information. In this, Harris linked up with another current in structuralism in viewing language as having a code-like character. We shall first need to briefly consider the widespread metaphoric employment in theoretical linguistics of terms from communication engineering, in particular, of 'code' and 'message' (and in particular that a 'code' carries the information or meaning of a 'message'). Subsequently, it will be shown how Harris's conception is at variance with much of this usage (e.g. Jakobson 1952: 559, 1961: 245–252).

As is generally known (for example, Mounin 1970: 141), Saussure had already introduced the term 'code' into structural linguistics by initially designating *langue* as the code utilized in the combinations of sound uttered in order to express the personal thought of the speaker (Saussure 1916: 31). In so doing, it is plausibly maintained that he was reviving, in rather explicit form, a traditional (and simplistic) psychological 'theory' of the relation between thought and language.[9] However, there seem to be two conceptually separate

9. One need only recall Bloomfield's ascerbic comment in an otherwise positive review: "Now de Saussure seems to have had no psychology at all beyond the crudest popular notions [. . .]" (Bloomfield 1923: 107).

rationales underlying his choice of this term. On the one hand, as is evident from the notorious 'circuit diagram' of speaker and auditor, Saussure (1916: 27–28) does seem to have naively assumed, at least for purposes of illustration, a traditional perspective on linguistic meaning and communication, the so-called 'translation theory of understanding' (Parkinson 1977 and Roy Harris 1987). On the other, the structuralist concept of the phoneme as a discrete, combinatorially treatable unit of sound — in Saussure's famous phrase, an entity 'wholly contrastive, relative and negative' (Saussure 1916: 164) — naturally suggested an analogy to telegraphic codes. It is worth pursuing each of these rationales in a little more detail, for, as we shall indicate, they are associated with different and conflicting views of the nature of language and the character of language structure.

The terminology of 'coding' and 'decoding' is but a relatively recent manifestation of an antiquated but powerful metaphor that still tends to dominate much thinking about linguistic communication and a fortiori about the character and nature of language structure. This is the image of a process of translation between 'ideas' or, in more *au courant* versions, some form of 'mental representation' and the external physical medium of language (either speech or writing). On this 'model', communication occurs when a 'message' or 'belief' existing in some (usually propositional) form of representation in the mind of one person ('the speaker') is 'coded' into a physical (acoustic, orthographic) form which is the medium for transmission to another person ('the hearer'). On receipt, the hearer then performs (usually unconsciously, of course) a 'decoding' of the message into a constituent set of ideas or mentally represented meanings. Accordingly, understanding between speaker and hearer occurs just in case the idea or mental representation elicited in the mind of the hearer is the same as, or sufficiently similar to, that originally in the mind of the speaker (e.g. Denes & Pinson 1963: 6; Fodor et al. 1974: 13–14).

The roots of this 'conduit metaphor' of communication (Reddy 1979) can be traced back to antiquity, to the Stoics and even earlier, but it was John Locke (building upon a doctrine of ideas common also to, e.g., the Port-Royal Logic) who gave it nearly canonical formulation in Bk. iii of his *Essay Concerning Human Understanding* (1690). Ecumenical in its appeal, it was adopted or tacitly assumed in both empiricist and rationalist traditions from epistemology and philosophy of language through philosophy of psychology. More recently, this doctrine, in its various and insignificantly different forms, the family of views which for obvious reasons may be termed 'mentalism', has enjoyed a

resurgence in linguistics and philosophy, despite prominent opposition from W.V.O. Quine and others (Quine 1960; Ziff 1960). In large measure, this has been due to an influx of ideas from communication engineering in the 1950s and 1960s, and to computer analogies from the 1950s onwards.[10]

On the other hand, and quite independently of the mentalist coding/decoding or translational 'model' of linguistic understanding, the supposition that the elements of *1angue* were solely discrete, contrastive, and only relationally individuated entities, led Saussure to repeatedly make the comparison of the system of language to telegraphic codes, such as Morse code (Saussure 1916:36), to the game of chess (43,125,149), and to an algebra (168). Later structuralism, particularly in America, placed considerable emphasis on the 'telegraph-code structure of language' without any accompanying commitment to underlying 'ideas' or mental processes of translation, indeed, while being antipathetic to talk of such processes. To some theoretically-minded structural linguists, the code analogy seemed especially suited to the characterization of language structure as comprising various levels of hierarchical constructions ('molecules') from, ultimately, a few dozen phonemes 'atoms').[11] But beyond this, Harris pointed out that in the discernible code-like properties of language structure lay the grounds of an explanatory account of its character. Observing that the elements of language are discrete, arbitrary, and preset by convention within a linguistic community,[12] while noting that such structural features are required for transmissibility without error compounding, Harris drew the conclusion that language structure is

10. Michael Dummett (1988:183–187) has recently identified the 'code conception of language' as lying behind the inversion of thought over language in the writings of Evans, Peacocke and others, an error, he claims, which threatens to lose Frege's hard-won gains against psychologism. Dummett argues, on grounds largely derived from considerations advanced by the later Wittgenstein, that recognition of the explicitly social character of language, and not Frege's 'mythology' of a 'third realm' of thought, stands at the center of the brief against psychologism in philosophy of language.

11. E.g., Joos (1950:705): "[Linguists] say, in effect, that the design of any language is essentially telegraphic — that the language has the structure of a telegraphic code, using molecular signals made up of invariant atoms, and differing, e.g., from the Morse code principally in two ways: the codes called 'languages' have numerous layers of complexity instead of only two, and in each layer there are severe limitations upon the combinations permitted."

12. It is this, Harris stresses, that makes the hearer's rendition of an utterance a repetition (not an imitation) of the speaker.

thereby also explicable as an instrument for transmission (not only communication) of information.[13]

From this second perspective, then, one can at most say that language is analogous to a code, not that language is a code. For a code is a 1–1 mapping between already well-formed expressions (the 'message') and the elements of a chosen cipher. Thus the grammatical structure of the message is presupposed.[14] But this assumption, equivalent to that of a metalanguage external to the language under analysis, is precisely one that should not be made in linguistics (Harris 1968:11, 1988:13, 1976:273 = 1981:389). For in attempting to specify how language 'carries' information, a linguist writing a grammar does not have the resources to reduce a given utterance or text to a prior 'message' inscribed in some 'internal language' or universal semantics or logical formalism. Such a reduction can only be carried out (if at all) by employing what is, from the point of view of grammar, an external metalanguage. But (as urged above) this is an illicitly privileged standpoint that masks the methodological importance of the social character of language: grammatical classification supervenes upon the shared behaviors of a speech community. Moreover, an external metalanguage is an idle posit that — being explanatorily circular — falls outside the domain of accountability of grammar.[15]

I have tried to establish that there are two distinct strands or affiliations for the prevalence of code metaphors in recent linguistics. Certainly under the influence of the terminology of communication theory, some prominent structural linguists, particularly in Europe (but including Jakobson in the USA), collapsed this distinction, speaking indifferently of the coding of thought in language and of the code-structure of language.[16] In several attempts to extend concepts from communication theory to phonemic analysis, the fact that no applicable non-statistical concept of information is developed in this theory has often been ignored or downplayed (Martinet 1964:172ff., and Malmberg 1967:

13. Harris (1968:6–8); more elaborately discussed in Harris (1988:87–113), and in Harris (1991), especially chapter 11.

14. This is also clearly pointed to by F. François, "Le langage," Encyclopédie de la Pleiade, Paris, 1968:11, cited by Mounin (1970: n. 8, 144–145).

15. This point is strikingly made in a different context in Friedman (1987:14–16).

16. For example, Benveniste (1966[1963]:30–31): "[. . .]on peut espérer des théories de l'information quelque clarté sur la manière dont la pensée est codée dans le langage." and 23: "la langue étant organisée systématiquement et fonctionnant selon les règles d'un code[. . .]".

Chapter 3). Harris proceeded in quite another direction. In *lieu* of assuming one or another notion of information and attempting to map the structural or 'logical' form of sentences into such a framework,[17] Harris showed that the (hierarchy of) restrictions on combinations of linguistic elements have a cumulative effect of *creating* information. Taken as an aggregate, they comprise what can be said to be the informational structure of the sentence, discourse, or sublanguage so characterized. In so doing, Harris showed that he was not only a theorist of language but also one of information.

4. Least Grammar

The informational character of language structure, coupled with the unavailability of an external metalanguage, leads at once to the methodological requirement of what Harris termed a 'least grammar'. In this, we again see how theory and method are inseparably connected. Given the association between redundancy and structure, it is obviously essential that the grammatical statement of each restriction be maximally efficient in the following sense: it must not contribute to the redundancies of combination of elements it seeks to describe. For grammatical inefficiency gives descriptive standing to what is only an artifact of method. Of relevance here is that the notion of redundancy, in this sense, can provide a means for distinguishing information in language. Somewhat like the situation in the theory of computational complexity (as developed by Solomonoff, Martin-Löf, Kolmogorov, Chaitin, and others), where the information of a string of digits in binary notation may be defined as a particular function of the length of the shortest program (also a string of digits) that computes it, the information of a text, set of discourses, or sublanguage, is expressed by the *minimal grammar that completely characterizes it in terms of its recurrent elements and their modes of combination.* Lacking an external metalanguage from which to derive these elements and their permitted combinations, the elements can only be set up purely distributionally, i.e., combinatorially.

---

17. Harris (1991:348): "We cannot in general impose our own categories of information upon language [. . .] We cannot determine in an a priori way the 'logical form' of all sentences [. . .] We certainly cannot map them in any regular and non-subjective way into any informational framework independently and arbitrarily chosen by us."

The statement of distributional relations is a presentation of linguistic observations, data displayed in a certain organization, e.g., a tabular arrangement. Distributional analysis was misunderstood and inappropriately criticized at the time and subsequently from both sides of the methodological fence. On the one hand, the young turks saw it as 'merely taxonomic' (as if an adequate classification of observations is unimportant) while the old guard termed it 'hocus-pocus linguistics' (as if increasingly general statement of structural relations was unnecessary). But in any case, mere statement of distributional restrictions was never intended as 'the goal' of Harris's structural theory. To the contrary, Harris's early and most detailed presentation of distributional methods concludes in a proposal that a grammar be presented as an axiomatic theory (Harris 1951a: 372–373, completed in 1946). Precisely what an axiomatic formulation of a grammar has to account for ('explain') is the observed range of co-occurrences of each element, its distribution in the language. The second requirement is incumbent upon the first. Since by definition, language structure is a structure of the restrictions on combinations of elements recognized by the speakers of the language in question, it is imperative that the characterization of this structure (in a 'grammar') not contribute to the redundancies of combination, the bearers of information in the language, to be described.[18] This is not just the general methodological virtue of economy of means. If language structure *is* informational structure, the requirement of a 'least grammar' is not a nicety, it is a necessity. Every restriction on combinations registered by the grammar must correspond to or correlate with a difference in information, a distinction intersubjectively attested to by speakers of the language. Through a process that Harris termed 'regularization' (Harris (1968) Chapter 6), the task of characterizing a language (or some restricted use of language in sublanguage or discourse) is that of replacing elements that have many apparent restrictions on their combination with less restricted elements that are recognized to 'say the same'. Even in his first book of 1951 this methodology is manifest in the attempt, through successive chapters, to continually seek more and more general classifications of linguistic elements (Harris (1951a), especially Chapters 7 through 19).

18. Harris (1982: 10–11): "[T]he grammatical description [must be kept] as unredundant as possible so that the essential redundancy of language, as an information-bearing system [. . .] not be masked by further redundancy in the description itself"; see also Harris (1968: 12fn16).

Subsequently, grammatical transformations are developed as a kind of 'extended morphophonemics', more powerful regularizing methods that enable even the derivation of tense and affixes. Throughout Harris's work, elements fall into the same equivalence class only if the individual members share a formally characterizable common environment of occurrence. On the hypothesis that each such formally stated environment is also recognizably distinct by speakers of the language,[19] grammatical methodology thereby acquires something of an 'operationalist' cast. But this stems not from outmoded scruples of positivist or behaviorist metascience, it is because all formal characterization of information-bearing elements in language requires empirical justification through correlations with speaker recognition of which elements are 'the same' and which are 'different'. Non-repetition is then a difference that makes a difference: this is to encounter language primarily as 'an item of culture'. Naturally, it is also because of the fundamentally social character of repetition that there can be no 'private language'.

5. A self-organizing system

I doubt that I am alone in marveling that in his last book on language and information (Harris (1991), chapters 11 and 12) Harris addressed the hoary philosophical question of how language 'connects with the world', here in the context of a *Gedankenexperiment* concerning the development of syntax. Characteristically, Harris offered a sketch of a constructive and developmental (one might also say, historical-materialist) answer that nowhere appeals, implicitly or explicitly, to a *deus ex machina*. We have, first, a pre-syntactic use of 'words' with primary referential meaning. Certain words (sound combinations) may be thought to have been said consequent upon the saying of others: e.g., *run* predicated of *deer*, *sleeps* of *boy* and *deer*, *red* and perhaps *sleeps* of *flower*, but not *sleeps* of *run*, or *flower* of *deer*. An understandable need for efficiency of communication, for minimizing ambiguity, conventionalizes and institutionalizes these differences in usage and eventually they are 'frozen out' as categorial differences. This is not to say that the concept of predication must have already been present. As Harris (1991: 369) notes with reference to Piaget,

19. Hoenigswald (1965: 192) gives an explicit statement of this hypothesis; see also Harris, (1954: 13).

one does not need to understand in order to do. But the predicational partial ordering of words — which creates syntax as a dependence on dependence relation among words — can plausibly be imagined to have emerged from these initial dependencies of one particular word on another. Once 'frozen out' as a difference in usage, this abstract pattern of relation among words can be readily extrapolated to express more and more complex sentences: e.g., *continue* said of *eats*, *swims*, but not of *boy*; *believe* said of *man* and *boy sleeps*, but not of *continue*, and so on. But this is to say that the syntax-creating partial ordering constraint emerges from the reasonable conjecture that certain words are initially said 'about' certain saliencies — i.e., objects and situations — in the perceived world. If one can speak here, as does Harris, of the 'co-occurrence' of certain properties and objects in a commonly perceived world, e.g., *red berry*, *red flower*, *large fish*, and so on, then one can understandably see in the constraints on word co-occurrences a reflection and recording of this experience, and no doubt very soon, even a substitute for it.[20] The partial-ordering constraint is sufficiently general to be readily extendible for expression of far more complicated types of predications beyond the simple attribution of salient properties to salient objects in a commonly perceived world. But these humble origins are easily obscured in the complexity of further emerging grammatical relations (in particular, in likelihood inequalities of word co-occurrences, and in processes of ellipsis, that is, reduction in phonemic shape). Further, even the concept of 'information about the world' may be constructively accounted for as arising from the meaning of the sentence-forming predicational constraint that itself is a conventionalization and institutionalization of directly referential word use (cf. Harris 1991:354). From this vantage point, the notorious difficulties in attempting to answer the question of how language 'hooks up to the world' stem from the non-developmental manner in which the question has been traditionally posed and treated. But from this perspective it is only obvious that form and content, syntax and semantics, must surely have developed hand in hand: in Harris's formulation, "content follows upon form" and "form follows upon content" (Harris 1991:354). So there is, after all, a kernel of truth in picture theories of meaning, but only that. This is not an endorsement of causal realism in philosophy of language; however, in locating the origin of grammatical relations in an initial purely referential and non-syntactical 'word'

20. Cf. Sapir, (1933:11): "It is important to realize that language may not only refer to experience or even mold, interpret and discover experience, but that it is also a substitute for it."

usage, an attenuated realism enters, as Harris once put it, "through the back door".[21] Considered as a system of representation, language does not, and cannot mirror reality; at best, it reflects, records, and transmits a salient order in a commonly perceived world. This perhaps was its original mission and, not surprisingly, this remains its primary function in research reports of an experimental science, as can be demonstrated in the information structures of the language of these reports.[22] Harris's hypothesis is that additional structures of language emerge as further conventionalizations of usage from this original, and primitive, referential function, producing constraints that only indirectly, or distantly, or not at all carry referential meaning. Within the additional latitudes of expression provided by new syntactical constructions are opened up new possibilities for purely symbolic or abstract vocabulary. Not incidentally, these further structures provide as well the wherewithal both for the development of abstract thinking, and for the formation of 'nonsense' (as opposed to ungrammatical) sentences. This is a highly plausible accounting of language as a self-organizing system developing in tandem with the complexity of thought.

6. Conclusion

Let us briefly consider the bearing of Harris's informational interpretation of language structure on the topic of explanation in linguistics.[23] There has been much discussion of the necessity for linguistics, as a science, to proffer explanatory theories as opposed to 'mere' descriptions of linguistic data or behavior. Influenced by prevalent realist currents in the philosophy of science, those engaged in a quest for an 'explanatory' linguistics have urged that the discernible regularities and patterns in linguistic data can only be accounted adequately explained, in the last analysis, by reference to underlying psychological and biological structures. According to this view, linguistics, with its reliance on socially and historically contingent linguistic data and with its 'abstract' characterizations of these underlying realities, is ultimately to be subsumed in some future science of the biology of cognition. For the time being, however, linguistics is to push ahead, seeking 'deeper' and more abstract theoretical character-

21. Harris used this metaphor in a conversation in 1986.
22. This claim is exhaustively documented, for papers written in cellular immunology in both English and French in the period 1935–1970, in Harris et al. (1989).
23. The topic of explanation in linguistics is extensively treated in Ryckman (1986).

izations, lying at farther and farther remove from the observable data of language. This is indeed an audacious program of research that appears to be fashioned, to a very considerable extent, on the model of the recent history of fundamental physics. Whatever the internal difficulties with such a program, it should be clear that the work surveyed here is of a completely different theoretical and explanatory orientation; as Harris observed, "generality is not the same thing as abstraction" (Harris 1981: v). My remarks have been concerned to show that in upholding the autonomy of linguistic theory (as manifest in the unavailability of an external metalanguage), Harris did not depend on outmoded strictures of positivist metascience, whether against unobservables or in favor of the instrumentalist character of scientific theories. Rather he pursued an explanatory account of language structure consonant with his view of language as a self-organizing system for transmission of information, evolving through a continual process of institutionalization of usage. When language is seen as a product of selective processes of social institutionalization, the uniqueness of its development is reduced, and the problem of language acquisition is correspondingly diminished. On this view, language is not, in any sense that matters to its function of communicating and transmitting information, located in the human genome but is a shared social practice *par excellence*, a point that seems especially appropriate if we consider the particular languages of the special sciences. Moreover, the very means by which language manages to 'carry' information is not something external to the language itself but is the structure of successive levels of constraints, each higher level presupposing those beneath, governing its elements. Language is the paramount human means, not of communicating meaning — for there are many non-linguistic or quasi-linguistic ways of doing that — but of articulating, delimiting, and transmitting meaning, as predication-structured information, between one individual and another or between a group and a wider community. It follows that the patterning existing in those social practices that we term a language is that of information.

References

Benveniste, Émil. 1963. "Coup d'oeil sur le developpement de la linguistique". In his *Problemes de linguistique générale*, Paris Gallimard, 1966, 30–31.

Bloomfield, Leonard. 1923. Review of Saussure (1916), second edition. In C. Hockett (ed.) *A Leonard Bloomfield Anthology*, Bloomington, Indiana University Press, 1970, 106–108.

Denes, Peter B. & Elliot N. Pinson. 1963. *The Speech Chain: The physics and biology of spoken language*. Bell Telephone Laboratories. Garden City, NY: Anchor Press.

Dummett, Michael. 1988. "The Origins of Analytic Philosophy", Part II. *Lingua e Stile* 23: 171–210.

Fodor, Jerry A., Bever, Thomas G., & Garrett, M. F. (1974). *The psychology of language: An introduction to psycholinguistics and generative grammar*. New York: McGraw-Hill.

Friedman, Michael. 1987. "Theoretical Explanation". In Richard Healy (ed.), *Reduction, Time and Reality: Studies in the Philosophy of the Natural Sciences*. New York & London: Cambridge University Press, 1–16.

Goldfarb, Warren. 1979. "Logic in the Twenties: the Nature of the Quantifier". *Journal of Symbolic Logic* 44: 351–368.

Harris, Roy. 1987. *Reading Saussure*. London: Duckworth.

Harris, Zellig S. 1951a. *Methods in Structural Linguistics*. (Reprinted as *Structural Linguistics*, "Phoenix Books".) Chicago: University of Chicago Press.

Harris, Zellig S. 1951b. "Sapir's Selected Writings", Review of Sapir (1949). *Language* 27.3: 288–333.

Harris, Zellig S. 1954. "Distributional Structure". *Word* 10.2/3: 146–162. Repr. in Harris (1970: 775–794) and in Harris (1981: 3–22).

Harris, Zellig S. 1968. *Mathematical Structures of Language*. (=Interscience Tracts in Pure and Applied Mathematics, 21.) New York: Interscience Publishers, John Wiley & Sons.

Harris, Zellig S. 1970. *Papers in Structural and Transformational Linguistics*. Ed. by Henry Hiż. [=Formal Linguistics Series, Vol. 1.] Dordrecht/ Holland: D. Reidel.

Harris, Zellig S. 1976. "On a Theory of Language". *Journal of Philosophy*, LXXIII: 253–276; partially reprinted in (Harris 1981: 377–391).

Harris, Zellig S. 1981 *Papers on Syntax*. Ed. by Henry Hiż. (=Synthese Language Library, 14.) Dordrecht: D. Reidel,

Harris, Zellig S. 1982. *A Grammar of English on Mathematical Principles*. New York: John Wiley & Sons.

Harris, Zellig S. 1988. *Language and Information*. (=Bampton Lectures in America, 28.) New York: Columbia University Press

Harris, Zellig S. 1991. *A Theory of Language and Information: A mathematical approach*. Oxford & New York: Clarendon Press.

Harris, Zellig S., M. Gottfried, T. Ryckman, P. Mattick, Jr., A. Daladier, T. N. Harris, & S. Harris. 1989. *The Form of Information in Science: Analysis of an immunology sublanguage*. Preface by Hilary Putnam. (=Boston Studies in the Philosophy of, Science, 104.) Dordrecht and Boston: Kluwer Academic Publishers.

Hoenigswald, Henry M. 1965. "Review of John Lyons, Structural Semantics". *Journal of Linguistics* 1: 191–196.

Hymes, Dell and John Fought. 1975. *American Structuralism*, The Hague: Mouton.

Jakobson, Roman. 1952 "Results of a Joint Conference of Anthropologists and Linguists". Supplement to *IJAL* 19.2: 11–21. Repr. in his *Selected Writings* 2, 1971, 554–567 (The Hague: Mouton).

Jakobson, Roman. 1961. "Linguistics and Communication Theory". *Proceedings of Symposia in Applied Mathematics*, vol. XII, Providence, R.I., American Mathematical Society, 245–252.

Joos, Martin. 1950. "Description of Language Design". *Journal of the Acoustical Society of America* 22:701–708.

Malmberg, Bertil. 1967. *Structural Linguistics and Human Communication*, 2nd revised edition, Berlin & New York: Springer Verlag.

Martinet, André. 1964. *Elements of General Linguistics*. Chicago: University of Chicago Press.

Mounin, Georges. 1970. "La Notion de Code en Linguistique." In *Linguistique contemporaire: Hommage A Eric Buyssens*. Bruxelles: Editions de L'Institut de Sociologie, Universite Libre de Bruxelles, 141–149.

Parkinson, G.H.R. 1977, "The Translation Theory of Understanding," in G. Vesey (ed.), *Communication and Understanding: Royal Institute of Philosophy Lectures*, vol. 10, 1975–1976, London, 1–19.

Quine, Willard van Orman. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Reddy, Michael. 1979. "The Conduit Metaphor: A Case of Frame Conflict in Our Language about Language". In Andrew Ortony (ed.), *Metaphor and Thought* (second edition 1993). New York & London: Cambridge University Press, 164–201.

Ricketts, Thomas G. 1985. "Frege, the Tractatus, and the Logocentric Predicament". *Noûs* 19:3–15.

Ryckman, Thomas A. 1986. *Grammar and Information: An Investigation in Linguistic Metatheory*. Ph. D. dissertation (Philosophy), Columbia University.

Sapir, Edward. 1929. "The Status of Linguistics as a Science". *Language* 5:207–214. Repr. in Sapir (1949:160–166).

Sapir, Edward. 1933. "Language". *Encyclopaedia of the Social Sciences* (New York: Macmillan) 9:155–169. Repr. in Sapir (1949:7–32).

Sapir, Edward. 1949. *Selected Writings of Edward Sapir in Language, Culture and Personality*. Ed. David G. Mandelbaum. Berkeley & Los Angeles: University of California Press.

Saussure, Ferdinand de. 1916. *Cours de linguistique generale*. C. Bally & A. Sechehaye (eds.). Paris/Lausanne: Payot.

Sheffer, Henry. 1926. "Review of *Principia Mathematica*, Volume I, second edition", *Isis* 8:226–231.

Simon, H.A. 1984. *The Sciences of the Artificial*. 2nd edition. Cambridge, Massachusetts: MIT Press.

Wittgenstein, Ludwig. 1922. *Tractatus Logico-Philosophicus*. London: Routledge & Kegan Paul

Ziff, Paul. 1960. *Semantic Analysis*. Ithaca: Cornell University Press.

Chapter 2

Some implications of Zellig Harris's work for the philosophy of science

Paul Mattick
Adelphi University

1. Language, philosophy, and science

For much of the twentieth century, the philosophy of science was preoccupied with problems of the nature of scientific language. Philosophy, seeking since the eighteenth century to justify secular claims to reliable knowledge, was naturally drawn to make sense of the claims to truth of a highly mathematicized physics operating conceptually and experimentally at a level far removed from what counts as 'observation' or 'experience' in the world outside of science. This situation was exacerbated by the revolutions in physical theory occasioned in the early 1900s by the advent of relativity and quantum mechanics. Given philosophy's traditional orientation to the analysis of 'judgments', linguistically expressed thoughts, the problem of the relation of scientific theory to experimental observation could easily take the form of that of the empirical significance of theoretical statements. The development of mathematical logic around the turn of the century by Gottlob Frege and Bertrand Russell suggested the possibility of analyzing theoretical discourse into truth-functions of sentences, on the one hand, and word meanings understood as mappings of concepts on experiences, on the other.

Something along these lines seemed to have been achieved in what was for quite a while the philosophical 'Received View' of science, that of logical empiricism. Assuming that logic provided the means for an analysis of language at a level more fundamental than that of ordinary grammatical categories, this view treated scientific theories as interpreted logical systems, in which the words whose substitution for variables made logical formulas into sentences acquired their meanings by their relation to elements of experience (represented variously as sensations, perceptions, or elementary sentences denoting perceptual

experiences). Thus the 'Received View' had two components, corresponding to the aim of explaining the relation of theory to experience and to the means of logical analysis. On the one hand, it distinguished between theoretical and observational terms; on the other, it analyzed word meaning into a conceptual 'sense' and the empirical 'reference' that it determined. According to this view, of which many forms descend from Frege's writings of the late 1800s, whereas syntax is a matter of purely formal relations between symbols, semantics is a relation of some sort between linguistic expressions and empirical entities such as physical observables, perceptual sensations, or facts. If theoretical expressions could be defined in terms of observational ones, and if the latter could be explained by reference to some sort of elements of experience, then the sought-for link between theory and observation would be established at the sentence level, as linguistic expressions would be shown to have meaning insofar as they can be observationally verified (or falsified).

However, no sooner was this program elaborated by Rudolf Carnap and others than the limitations of logic as a framework for analysis of the language of science began to emerge, for what one might call internal and external reasons (for an important early collection of papers on the collapse of the Received View, see Suppe 1977). Internally, the attempt to formalize the 'logical syntax' of science led to the disclosure of fundamental problems with the logical construal of laws and with the concept of confirmation. In addition, even as followers of logical empiricism's foundationalist approach were attempting to analyze scientific language in terms of truth-functions of elementary sentences, the assumption that science discourse (not to mention natural language as a whole) could be shown to be importantly characterized as truth-functional began to look doubtful for many reasons, such as problems with the analysis of counterfactuals. No less important a figure than Ludwig Wittgenstein, whose earlier writing was a crucial source for positivist theory of language, attacked linguistic along with mathematical logicism: Mathematics, he wrote in opposition to the views of Frege and Russell, no more needs a foundation 'than propositions about physical objects–or about sense impressions–need an analysis. What mathematical propositions do stand in need of is a clarification of their grammar, just as do those other propositions' (Wittgenstein 1964: 171e).

Meanwhile, the theory of meaning that identified the empirical significance of an expression with its experientially determined referent led to unpalatable difficulties. The most notorious of these was the thesis of the incommensurability of theories. This thesis seemed an inescapable consequence of the logical empiricist approach to meaning as soon as it became

clear that scientific statements were not accepted or rejected singly but as embedded in systems of concepts and principles. If no individual sentence could be confirmed or disconfirmed by itself, but only in connection with other statements, then experience determined meaning not sentence by sentence, and a fortiori not word by word, but only by way of the role of an expression in the language-carried conceptual system as a whole. It follows from this, however, that what might look like the same concepts — for instance, 'mass' in classical and relativistic physics — actually has not only different senses but different reference in the two theories. Hence no observation of 'mass' could be simultaneously relevant to both theories and so help decide between them: the theories would be 'incommensurable'. As Nancy Nersessian pointed out, in her excellent survey of this development, this wholly implausible conclusion established the failure of the classical (Frege-style) theory of meaning, so that "we need to construct a new theory of meaning for scientific theories" (Nersessian 1984: 23).

In association with these developments, an external critique of logical empiricism emerged from a growing interest in the historical development of scientific theories, as opposed to their analysis as finished products. The social-historically informed work done in France by Georges Canguilhem and his students, especially Michel Foucault, paid close attention to discourse without the assumptions of logical structure or empirical foundations for meaning. Within mainstream anglophone science studies, too, it came to be felt more strongly, as Dudley Shapere put it

> that what is needed is a closer examination of actual scientific development and practice — of the jobs performed by terms and statements in their actual employment in science, and of the respects in which those jobs change and remain the same as science develops. (Shapere 1984: 29)

But how, in what terms, is such an analysis to be carried out? In particular, there seems to be no substitute for one strength of the logical empiricist tradition, the formalization of meaning relationships and patterns of argumentation. It appears that the only alternative is careful examination of science on the basis of the philosophical investigator's semantic intuitions with respect to 'the jobs performed by terms and statements' and their changes over time. Good examples are Shapere's own study of the use of 'observation' (along with 'experiment' and 'detection') in astrophysics, and his investigation of the development of nomenclature and description in chemistry (Shapere 1982, 1984: Ch. 15).

Shapere's approach led him, for instance, to the promising concept of the 'domain' — a body of information bearing on a particular problem and utilizing characteristic methods of study–as a substitute for the earlier conception of a 'theory' as a formal system interpreted by linked theoretical and observational sentences (as well as for the ill-defined concept of 'paradigm'; see Shapere 1984: chs. 13, 14). But the promise was not kept; the terminology has not been developed and no serious replacement for the logical empiricists' 'theory' has been established. 'Domain', in contrast to 'theory' understood as a matter of axiomatic systems, lacks precise definition in terms of the units of information and the relations between them that create the basis for scientific work. (The related concept of 'interfield theory', intended to explain the relation between scientific fields or domains, is necessarily even less precisely defined [see Darden & Mull 1977: 45].) But precision is desirable above all in questions about the nature of science employing concepts like 'meaning' and 'theory change'. How are we to distinguish or identify as 'the same' various uses of key scientific terms, and so track the development of scientific accounts of the world? Without the logical empiricist model of theories how are we to identify the objects of study in the philosophy of science; to use Shapere's term, how are we to define the boundaries of a domain, establish the relations between domains, or discuss the relations between methods, concepts, and observations within a domain?

Nersessian concluded, from her survey of the fate of logical empiricist syntax and semantics, that 'the 'linguistic turn' in philosophy led philosophy of science away from an examination of its subject — science' (Nersessian 1984: 28). However, as she did not note, it also led philosophy away from an examination of the language of science as an aspect of socalled natural language. In disregarding this Nersessian is not alone; in general, philosophers of science have ignored the power of linguistic analysis to illuminate issues in which they are interested. Yehoshua Bar-Hillel was unusual in taking the trouble (in 1954) to deny that distributional methods had much to contribute to the understanding of linguistic information and in particular to argue against Z. S. Harris's claim of the power of such methods for the analysis of semantic relations. Bar-Hillel asserted, following Carnap, that logical concepts could be defined in purely syntactic terms. But this syntax could not be identified with the grammatical relations discernable by distributional analysis (Bar-Hillel 1964). In line with the Received View, Bar-Hillel assumed that (analogues of) logical properties and relations can be unambiguously identified for sentences in natural language, so that inter-sentence relations gener-

ally, or at least in science language, can be usefully analyzed in logical terms — an assumption that had to be given up as soon as the non-atomic character of science sentences became apparent, together with the importance of non-truth-functional relations between them (for a rejection of natural language logicism on the basis of linguistic data, see Harris 1991:305 ff.).

In fact, as Harris observed

> Not a few of the difficulties in the philosophy of language and in neighboring areas of philosophy arise from starting with the equipment which had been developed for truth systems and using it to analyze the information system that language represents. (Harris 1976b:252)

This observation is drawn from one of a pair of articles in which Harris made the philosophical relevance of his linguistic investigations available in philosopher-friendly form as early as 1976 (Harris 1976a, b). That these articles have been, so far as I am aware, completely ignored by philosophers of language and science reflects the hold on them of disciplinary tradition, which has, for instance, made it difficult for analytic philosophers to abandon the Fregean syntax/semantics distinction and the supposed centrality of logic to an understanding of linguistic information.

Expressing the postpositivist movement away from formalization, Nersessian rightly says that

> the nature of meaning in scientific theories must be seen in the context of the network of beliefs (theoretical, methodological, metaphysical, common sense) and problems (theoretical, experimental, metaphysical) which is part of the making of meaning in scientific practice [. . .] (29).

These beliefs and problems, as she is aware, are available for study as formulated in language (together with such other symbolic modes, in some cases, as diagrams and pictures). Of course, 'scientific practice' cannot be reduced to discourse. But, as Harris's work demonstrates, ignoring the study of science language has meant neglecting important means to the understanding of the nature of science.

2. Language and Sublanguage

What makes a philosophically-relevant analysis of science language possible is the fact that language has a discoverable structure which, while significantly different from that of logic, may be given a compact formal description. That

is, 'It is possible to formulate an abstract system, i.e. one whose objects are defined purely by the relations in that system, which is adequate precisely for natural language' (Harris 1968: 176). While this abstract system is in the first place formulated in purely combinatorial terms, its elements in their relations can be seen to have clear semantic properties. In particular, all occurrences of natural language can be described as word-sequences satisfying certain combinatory constraints, and these constraints can be given an informational interpretation. The first constraint, on word entry into sentences, consists in a partial ordering on words, such that the occurrence of a word (an operator) depends on the occurrence of another set of words (its argument[s]). This ordering partitions the set of words into a small number of operator/argument classes. Thus, words like *drink* occur only together with pairs of words like *boys*, *milk*; words like *believe* only with argument pairs like *mothers*, *drink*; so that this constraint permits sentences like *Mothers believe boys drink milk* and excludes sentences like *Boys believe milk* and *Mothers drink boys believe milk*. The semantic relation carried by this constraint is (loosely) that of predication: In these examples, drinking is predicated of boys with respect to milk and belief of mothers with respect to boys' drinking of milk.

A second constraint accounts for a more detailed level of sentence meaning, as it consists in the varying likelihood with which, for a given argument, different words within its operator class appear with it (thus *drink* is more likely than *eat*, which is more likely than *build*, as operator on *boys*, *milk*). This constraint is one important feature differentiating language from formal logic: While the latter maps well-formed formulas onto the two values True and False, the former places well-formed sentences (i.e., sentences meeting the operator-argument requirements) on a many-valued scale of acceptabilities (Harris 1968: 203). The classical syntax-semantics distinction — taken for granted, notably, by Harris's student Noam Chomsky in *Syntactic Structures* (see Chomsky 1957: 15) and still, despite apparent problems with it, in *Aspects of the Theory of Syntax* (Chomsky 1965: 16; see also 148 ff.) — breaks down, since every well-formed sentence, however unlikely its occurrence (even *Colorless green ideas sleep furiously*) has meaning in an appropriate linguistic context. In fact, the distinction between syntax, characterizing well-formedness in an object language, and semantics, necessarily formulated in a metalanguage, is inapplicable to natural language which, as Harris stressed, has no external metalanguage.

A third constraint on sentences regulates the reduction of their length by changing the shape (to zero, for instance, or to affix form) of words that

contribute little information in a particular verbal environment. For example, *Boys drink milk and girls drink milk* can be reduced to *Boys and girls drink milk* (while reduction of *Boys drink milk and boys eat cookies* to *Boys drink milk and cookies* is excluded). The semantic content of these reductions is exactly the redundancy of the material dropped or otherwise shortened (as when a pronoun indicates cross-reference to an object already mentioned). The third constraint thus brings in a different aspect of meaning. The first two constraints so to speak construct a statement by choosing among the possible predicates of a word another word with a particular likelihood of co-occurrence. The third constraint regulates not the formation of statements but changes in their shape. In fact, a small number of types of such reductions, which included the sentence transformations studied by 'generative grammar', suffice to derive all known sentence forms from unreduced sentences satisfying the first two constraints (Harris 1968, 1982). They therefore can be said to produce paraphrases of sentences formed by the two constraints.

The fact that only redundant material can be reduced means that information — in the standard sense — is not lost in reduction. Indeed, we can now identify linguistic 'information', for present purposes, as the semantic content of a sentence that remains constant under paraphrastic transformations. Information then appears as the product of the first two constraints and is unaffected by the changes regulated by the third.

It is interesting to note the light that this discovery of Harris's sheds on the distinction logical empiricism tried to draw between the assertion of empirical content characteristic of science and other uses of language, e.g. literary or metaphysical, in which meaning is not reducible to information but involves such features as literary form. It is of course not true that we cannot identify rhetorical devices characteristic of scientific writing. For example, observations and laboratory reports are commonly written in the passive voice (to emphasize the data gathered rather than the scientific worker gathering it), and there is a clear bias towards concision of expression. However, paraphrase and even translation are not seen as altering the content of science discourse. The problem with the logical empiricist attempt to demarcate a region of 'empirical significance', identified with science sentences, within language was the definition of this region in terms of a philosophical conception of 'experience' that in the event proved unworkable. In contrast, Harris's distinction between information and paraphrastic transformations of it is made in terms of inspectable features of language itself, without making any philosophical claims about the relation of language to a represented 'reality'.

By undoing reductions the sentences of a text or group of texts can be put in a directly comparable form without alteration of information. This is not logical form: the operator-argument structure is not that of predicate calculus nor are higher-order operators truth-functional. It does, however, facilitate the analysis of connected discourse. What makes such analysis pursuable to a high degree of detail is the fact that a body of sentences in texts grouped by subject-matter may show constraints on word co-occurrence more specific than those holding for the language as a whole.

It is not just that particular areas have restricted vocabularies, but that only particular types of statements can be made in terms of such vocabularies. In immunology, for instance, one can say *Cells produce antibody* but not *Antibody produces cells*, although the latter is a possible sentence in English as a whole. Analysis of sentences in a given field, that is, shows that they are constructed of just certain word subclasses, which are combined (under the operator-argument constraints) to form a rather small set of sentence types. We can call these sentence types 'information structures' to indicate that they provide an inventory of the kinds of facts constituting the information provided by texts in the field.

These sentence types or information structures are so restricted, as sequences of word categories, as to make possible description of the texts categorized by them as instantiating a bounded sublanguage, i.e., a set of sentence types closed under linguistic operations. As the use of the term 'sublanguage' may be taken to signal, we return here to aspects of Carnap's program in *The Logical Syntax of Language* (as also to W.V.O. Quine's reformulation of that program in *Word and Object*): formal representation of scientific language, with a canonical form revealing the informational structure of sentences and of their interrelation in connected discourse, in such a way as to facilitate study of the semantic and reasoning relations between linguistic units. Because the form of the representation is not a logical one, into which the natural-language material of science must be fitted, but is that of the syntax of language itself, patterns of word use, structures of argumentation, and even the constitution of scientific fields and their interrelations can be studied directly at given points of time and in the process of change. On the other hand, since the words (and so the information structures into which they enter) are identified not by their meanings but by their combinations with other words, the method of analysis does not depend on the investigator's semantic intuitions so that it is replicable and is in principle mechanizable.

3. A case study: immunology

The price to be paid, of course, is detailed analysis of a sizeable corpus of discourse; the labor involved is more that of scientific work itself than the traditional philosophical procedure of reflection on some studied subject area. Harris et al. (1989), *The Form of Information in Science: Analysis of an immunology sublanguage*, demonstrates what can be done along these lines. The object of inquiry was cellular immunology during the period 1935–1970; the articles were selected in consultation with workers in immunology as representative of the discussion in that field during this time, which centered on the question of the cellular source of antibody. While this area of biomedicine is less mathematicized than the physical theories of primary interest to twentieth-century philosophy of science, it is certainly true that, to cite the words of a recent textbook, immunology is 'an entirely typical science whose history presents a standard case of how knowledge and experimentation evolve within a given domain of scientific inquiry' (Robert Klee 1997: 8; it is striking that this work, devoted entirely to immunology, betrays no knowledge of the existence of Harris et al.1989).

The first step of the study was a grammatical analysis of each sentence of the corpus of articles. Paraphrastic transformations — both the undoing of reductions and the utilization of alternative sentence forms, such as passive for active or the reverse — were employed to put the sentences into a directly comparable (canonical) form. This made possible the next step, the formation of classes of words having the same (operator-argument) grammatical relation to particular other words. For example, *Lymph nodes contain antibody* would be transformed into *Antibody is contained in lymph nodes* in order to present it as a variant of *Antibody is found in the lymph nodes* with *contained in* and *found in* (along with *produced in* and others) members of one word class, *V* (defined by the environment *antibodies . . . lymph nodes*). In the same way, we can identify a class of words *A* (including *antibodies*, *agglutinins*, and others) and a class *T* (*lymph nodes*, *lymph*, *serum*), which are arguments (respectively subject and object, in conventional grammatical terms) of the operators in *V*.

Since the word classes are defined in terms of word co-occurrences, we derive at the same time a sentence structure *AVT*. Another sentence type common to many of the articles is *GJ* (*G*: *antigen*, *bacteria*, *typhoid vaccine*; *J*: *inject*, *incision*, *introduce*, *vaccinate*); the two often appear under an operator like *after*, *following*, etc. (which fall into a class ;), as in *Following injection of antigen*, *antibodies are found in the lymph nodes*. This sentence as a whole can

therefore be given the representation *GJ*;*AVT* (with a shift in position of the ; class for representational convenience). The result of this analysis is a rather small set of word classes and sentence types constructed out of them sufficient for the representation of the corpus of texts, in the sense that every sentence of the articles can be mapped (after transformational decomposition) onto these sentence types. In terms of these word classes and sentence types, that is, we can formulate the syntax of the sublanguage of immunology.

The sublanguage classes do not correspond to word classes of English, but are rather superclasses of the latter. This reflects the fact that the analysis need only be detailed enough to reveal those sentence types peculiar to the sublanguage, and it also indicates that while every sublanguage sentence is a sentence of some natural language, and is therefore described by the grammar of the language, we can characterize a sublanguage by a grammar distinct from that of the whole language although related to it in stateable ways (see Kittredge & Lehrberger 1982: chs. 3, 4, 11). The word classes of this grammar may be taken as giving the types of objects, and the sentence formulas as stating the relations between these objects, with which the field is concerned. Thus a statement of the syntax of the science is at the same time a representation of the information which is its content.

This means that, given the definitions of the word classes (and, importantly, subclasses), the content of the corpus of articles can be represented by a sequence of sentence formulas. An important confirmation of the validity of this linguistic approach to information is the possibility of identifying, solely by reference to these (syntactically-defined) word-class formulas, those statements in articles that constituted major theoretical turning-points in the development of the field; that is, 'differences in word classes and in sentence formulas appear where there are known differences in information or in opinion' (Harris et al. 1989:64).

There are large portions of the texts of the articles investigated that are not structured by these word classes and sentence types. However, these portions are related to the sentences so structured in specifiable ways. An important example is that of 'metascience' portions of sentences, in which something is said about a statement proper to the science itself. (The metascience operators, that is, are like *believe* in *Mothers believe boys drink milk*.) The structure of this material cannot be analyzed into sequences of sublanguage word classes but it has a particular grammatical relation to material that can be so analyzed: it consists of (a) operators on material represented by sublanguage sentence types, and (b) first arguments of those operators that are not identical with the

subjects of the sentences that are their second arguments. Thus, in *The direct demonstration by McMaster and Hudack of the production of antibody in lymph nodes*, the phrase *the production of antibody in lymph nodes* (as a transform of *antibody is produced in lymph nodes*) is a case of *AVT* and *demonstration* — with its subject derived by transformation from *McMaster and Hudack demonstrated*, itself under the higher operator *is direct* — is said about that science sublanguage sentence. The class *M* of metascience operators includes such words as *find*, *study*, *investigate*, *report*, *believe*, *know*, *hold*, *contend*, etc. While clearly necessary to discourse in the field of immunology, they are not peculiar to it; conversely, distinguishing this class of material makes possible a definition of (immunology) science sentences as those sentences (after transformational recomposition of reduced sentences) that are arguments of *M* words. And these sentences are those that are mappable onto the sublanguage formulas.

One important result of the investigation, then, is that a class of sentential material distinct from, and said about, the information specific to the particular science occupies a syntactically specific position in relation to it. Distinct subclasses of this material can be distinguished, by way of their co-occurrence relations with the science information sentences, which are used to state relations between the materials, methods, and results of scientific work.

Thus there are operators, such as *show* in *The data presented show that following an injection of virus there is a burst of activity of the lymphatic system*, which can be thought of as stating 'evidential' relations between groups of sentences (others in this group are *report*, *describe*, *establish*, *suggest*, *confirm*). Another group of words — including, on the one hand, words like *data*, *results*, *observations* and, on the other, words like *methods*, *approach*, *procedure*, *technique* — classify science-information sentences into groups. Such words are used not only to state relations between subsets of sentences but also to impose a structure on the group of sentences constituting an article, seen most clearly in the labeling of article sections as Methods, Results, Discussion.

While these (and other) subclasses of metascience expressions can be distinguished, these classes do not show regular patterns of co-occurrence with the other word classes and sentence types of the immunology sublanguage. They can therefore be considered to be expressions in English, which are integrated into the immunology sublanguage (itself, of course, a subset of English; though see below) by being given the particular syntactic position of operators on immunology sentences. Similarly, there is a group of conjunctional expressions like *cause*, *is due to*, *is related to*, *accounts for*, *is associated with*, etc. These operators on pairs of science sentences must also be consid-

ered expressions in English (or 'scientific English') used to state relations between scientific facts. Such a relation may even be incorporated into the sublanguage grammar to the extent that a relation between facts acquires the status of a fact itself, as with the ; operator described above, which asserts the (temporal) linkage of *entrance of antigen* and *appearance of antibody* as the basic phenomenon to be understood.

This situation is different with another group of expressions that cannot be analyzed into sequences of sublanguage classes. These are the sentences (such as *Typhoid bacilli were washed three times*, *Spleen cells were centrifuged*) found primarily in the 'Materials' and 'Methods' sections of articles. These may well form part of a 'methods sublanguage', which would be common to a number of related fields. If this turns out to be true, we would have here an integration of external material into the immunology sublanguage by way of certain word classes being in the intersection of two sublanguages; thus *antigen* brings contexts like *was injected into rabbits*, found in the Methods material, into the immunology sublanguage. In the same way, results obtained in other, distinct areas of study (such as genetics or hormone chemistry) are incorporated into immunological discourse. A related case is that of arithmetical statements, which form a closed sublanguage and play an important role in immunological research.

While the metascience material can be seen as forming a level of language syntactically higher than that of the science sentences (in the sense of operating on the latter), the kinds of material just discussed constitute distinct though intersecting sets of sentence types on the same (syntactic) level as that of the immunology sublanguage. This corresponds well to our intuitive understanding of 'methods and materials' sentences (and likewise sentences in the 'Results' and 'Conclusions' sections of articles) as reporting facts, in this case about scientists' activities in the laboratory, while the metascience expressions state relations holding between these various sorts of facts (see Harris et al. 1989: ch. 6).

4. Philosophical Implications

While it is obvious that only fragmentary and tentative conclusions can be drawn from a single experiment with a novel method (novel for these purposes), philosophically interesting aspects of scientific discourse do emerge from the immunology study.

(1) To begin with, the world presented in this corpus is, to use Wittgenstein's phrase, a world of facts and not of things: first-order arguments (roughly, nouns) occur only under operators. We can distinguish a set of facts peculiar to immunological discussion as stateable in a bounded sublanguage; these are combined with each other, and with sentences describing materials and methods (including measurements and calculations), into more complex groups.

The fact complexes are only in part formed by the truth-functional operators *not*, *or*, *and*. In addition to various contrastive conjunctions (e.g. *however*, *nevertheless*), whose meaning is not fully rendered by translation into logical *and*, and argument-tracing ones (*so*, *thus*, etc.) distinct from logical *if . . . then*, there is a large group of operators stating correlational and causal relations between facts, which seems crucial to scientific argument. There are also modal operators (ranging from *is a fact* through *is possible* to *is very likely*), with a range of informational effects not formalized by the 'possible/necessary' dichotomy of current modal logics. The metascience operators are also not truth-functional, since they determine what philosophers call 'oblique contexts'. All this reflects the fact that, although logical structures are to be identified in natural language discourse, as in all cognitive activity, natural language is a much richer system of representation than logical calculi. The question of the relation of logic to the other sorts of structures involved in scientific work must await a more thorough analysis of the types of discourse structures — in particular, forms of argument, for instance, from experimental evidence and theoretical premises to proposed conclusions — to be found in science.

On the other hand, the work done already suggests an interesting approach to the problem as to what sort of entities the variables in logical sentences are to be taken as ranging over — as Susan Haack has put it, the problem of 'the relation between formal and informal arguments: what in informal argument corresponds to the well-formed formulae of formal languages?' (Haack 1978: 74). 'Statements' and 'propositions' are familiar answers, where these are understood as the meaning-content shared by some set of synonymous declarative sentences. Equally familiar are the problems that arise in trying to formulate a precise criterion for sameness of statement for distinct utterances.

To this philosophical difficulty Harris's approach offers not a solution but two directions in which the problem can be interestingly reoriented. First, the distinction between paraphrastic reductions and the basic sentence-forming constraints allows us to define a concept of information relevant to the

identification of statements or propositions, taken as informationally-equivalent sentences, and even permitting discussion of degrees of synonymy (see Harris 1991: 329 ff.). This gives an otherwise intuitive concept like 'statement' some operational rigor. Second, with respect to the specifically logical character of propositions as truth-bearers, that is, as sentences characterized as true or false, it is to be noted that science sublanguages present a situation at variance from that of language as a whole. As noted earlier, science discourse, in contrast to ordinary language, presents sharp restrictions on word co-occurrence. In biochemistry, for example, one can say, *The polypeptides were washed in hydrochloric acid*, but *Hydrochloric acid was washed in polypeptides*, while a grammatical English sentence, cannot appear in a biochemistry article. For the sentence-types peculiar to the sublanguage, the acceptability grading that is characteristic of natural language is tendentially replaced by an assertibility/nonassertibility dichotomy, where assertible sentences are those which can be either true or false. Science sentences, that is, have the truth-bearer characteristic of logical sentence variables.

Other features emerging from the immunology language study are relevant to this matter. Analysis of the immunology sublanguage in French as well as in English showed that they were in all essentials identical, sharing the same word classes and sentence types. This suggests that articles in whatever language in the field can be represented by sequences of the same formulas. This is related to the fact that formulas free the representation of information from noninformational features of language, such as the difference between active and passive. The science sublanguage formulas dispense with everything except what is relevant to the information distinguishable in the given field. It is therefore not surprising that the same formulas represent the same information irrespective of the language used. Here we have, therefore, something like the tenselessness of logical truth and the independence of logically relevant information from linguistic specificity. We can thus make sense of the particular relevance of logical reasoning to scientific inquiry, even while recognizing its limits in the face of the greater informational richness of language.

(2) While there is a clear distinction between scientific and metascience material, formulable in syntactic terms, it is clear that the latter, while distinguishable from the science sublanguage proper, is not outside science, nor even outside the subscience of immunology, but plays a fundamental role in articulating and organizing the information specific to or relevant to that subscience, by providing classifiers and argument structures (see Michael Gottfried, Chapter 5 in the present volume).

(3) This is related to the question of the determination of the boundaries of scientific domains or fields and of their relations to neighboring domains, background information, interfield theories, and the like. A science sublanguage grammar can be thought of as characterizing the texts reporting work in a socially recognized area of investigation. If we identify the sentences of the sublanguage as constituting the 'core' science language of (in this case) immunology, we can assign definite roles to material entering immunology texts from other sublanguages or from the language as a whole. In this way facts constituting a domain can be shown to be related to each other: by sharing the information structures peculiar to the domain (its 'core'); by being asserted, though not in the core sublanguage, as part of the same texts as 'core' sentences, or even about immunological objects (as in 'methods' sentences); or by more general (metascience) structures representing information about experimental and theoretical manipulations.

As Robert Klee explains in his *Introduction to the Philosophy of Science*, in the course of demonstrating the inadequacy of Popperian falsificationism as a picture of scientific practice, 'no general hypothesis by itself implies anything about an actual particular state of affairs', but only that hypothesis in the company of background information. 'What implies an observational prediction is that theory together with a myriad of interdependent beliefs, presumptions, guesses, and other theories' (Klee 1997: 72). The idea of implication by guesses suggests the inherent weakness of exclusively logical analysis of this situation; more important for present purposes is that science discourse analysis makes the role of background information visible in a form amenable to further detailed investigation.

(4) The 'syntax of science', so understood, suggests also an approach to another central concern of the philosophy of science, the analysis of meaning change in the development of scientific domains. While it is more than likely that there are important differences in this regard between different sciences, we can show in the case of immunology how, for example, the reclassification of plasma cells and lymphocytes — each of which was originally advanced as the source of antibody — as developmental stages in one cell type was accomplished in such a way as to leave undisturbed the reference of the two terms (since the cells were defined by features) and the contexts of their use (except those touching on the relations between them). Thus a theoretical conflict was resolved without loss of information, by preserving existing sentence types, and with the addition of new information (in the form of a new sentence type, *Cell X develops into cell Y*). This shows more clearly than ever that science

develops not by the accumulation of facts but by their redefinition and interrelation into structured bodies of information.

I am not arguing here either that all important problems in the philosophy of science can be resolved by the application of Harris's methods, or even that these methods should supplant other modes of enquiry into the issues discussed above. For one thing, these methods are extremely costly to apply, if only in terms of time. Nonetheless, philosophers of science would do well to investigate their utility for the investigations of many problems of interest, such as the actual structures of argument in science, and their relationship to structures of logical validity; the relations between experimental observation and theoretical conclusions, to be discovered in the particular syntactic structuring of Methods and Conclusions sections of articles; the nature of reference under conditions of theory change, and the implications for conceptions of scientific progress; the differences between more and less deeply mathematicized sciences, and the relations between mathematically- and linguistically-expressed areas of theory in the same science.

Finally, they are relevant to a question lying at the origin of the philosophy of science, the intuition that 'science' refers to a distinguishable area of practical and intellectual activity. Is this, as some suggest, merely an aspect of an ideology of modern culture, or does science involve particular methods and styles of information production and transformation that differentiate it from other modes of discourse? Here again the detailed analysis of discourse Harris's work has made possible offers the possibility of deeper understanding.

For instance, the sublanguage property is not unique to science, but is shared by other uses of language in (relatively) closed domains of information, such as law, technical manuals, etc. It is therefore not sufficient for the identification of scientific discourse. Other features — characteristic patterns of discourse structure, for example linking observation reports to theoretical hypotheses — which are not yet well understood, are extremely important. On the other hand, the sublanguage property is not necessary either, since one cannot deny the name of science to certain fields — for example, the critique of political economy, and perhaps areas in biology — that lack this type of highly restricted syntax. We can, however, at least explain the intuition that the socalled hard sciences represent central features of the concept of science, by noting their possession of the sublanguage property. We might speak of a continuum, with fields like history and anthropology — and perhaps even literary criticism and parts of philosophy — at one end, closer to the relatively unrestricted combinability of natural language, and fields like plate tectonics,

microphysics, and immunology, in which sublanguages can be identified, at the other end. Formulated in these terms, a science of discourse would have much to contribute to the understanding not only of science but of human practices generally.

References

Bar-Hillel, Yehoshua 1954. "Logical Syntax and Semantics". In *Language and Information: Selected essays on their theory and application.* Reading: Addison-Wesley.

Carnap, Rudolph. 1959. *The Logical Syntax of Language.* Tr. by Amethe Smeaton. Patterson: Littlefield, Adams, and Co.

Darden, L. & Maull, N. 1977. "Interfield Theories". *Philosophy of Science* 44: 43–64.

Haack, Susan. 1978. *Philosophy of Logics.* Cambridge: Cambridge University Press.

Harris, Zellig S. 1968. *Mathematical Structures of Language.* New York: Wiley-Interscience.

Harris, Zellig S. 1976a. "On a Theory of Language". *Journal of Philosophy* 73: 253–276.

Harris, Zellig S. 1976b. "A Theory of Language Structure". *American Philosophical Quarterly* 13: 237–255.

Harris, Zellig S. 1982. *A Grammar of English on Mathematical Principles.* New York: Wiley-Interscience.

Harris, Zellig S. 1991. *A Theory of Language and Information: A mathematical approach.* Oxford: Clarendon Press.

Harris, Zellig S., M. Gottfried, T. Ryckman, P. Mattick, Jr., A. Daladier, T. N. Harris, & S. Harris. 1989. *The Form of Information in Science: Analysis of an immunology sublanguage.* Boston Studies in the Philosophy of Science 104. Dordrecht: Reidel.

Kittredge, R. & J. Lehrberger (eds.) 1982. *Sublanguage: Studies on language in restricted semantic domains.* Berlin/New York: De Gruyter.

Klee, Robert 1997. *Introduction to the Philosophy of Science.* New York: Oxford University Press.

Nersessian, Nancy 1984. *Faraday to Einstein: Constructing meaning in scientific theories.* Dordrecht: Martinus Nijhoff.

Quine, W. V. O. 1960. *Word and Object.* New York: Technology Press & John Wiley.

Shapere, Dudley 1982. "The Concept of Observation in Science and Philosophy". *Philosophy of Science* 49: 485–525.

Shapere, Dudley 1984. *Reason and the Search for Knowledge.* Dordrecht: Reidel.

Suppe, Frederick 1977. *The Structure of Scientific Theories.* Second ed. Urbana: University of Illinois Press.

Wittgenstein, Ludwig 1964. *Remarks on the Foundations of Mathematics.* Tr. by G. E. M. Anscombe. Oxford: Basil Blackwell.

Chapter 3

Consequences of the metalanguage being included in the language

Maurice Gross
Laboratoire de Linguistique Informatique, NRS-Université Paris Nord

On several occasions, Z. S. Harris stated that the metalanguage of grammar was part of the language. At first sight, this statement is disturbing, but when understood in respect to Harris's practice of grammar construction, it has far-reaching consequences. In principle, the metalanguage of a scientific field is made of concepts and of statements involving these concepts: the laws of the field. In quantum physics for example, concepts are elementary particles, Planck's constant, etc., and statements are Heisenberg's uncertainty relations, etc. In syntax the concepts are essentially the grammatical categories of words (i.e. the parts of speech), and statements are the rules that assemble the words and/or categories into higher units such as phrases and sentences. Modern structural linguists, such as Leonard Bloomfield,[1] set out to formalize the metalanguage, and this activity has become the main trend, whether in generative syntax or in the various logical systems that aim at representing meaning. Meanwhile, the corresponding descriptive work has all but disappeared, at least for languages such as English that should be the main empirical background for theories. Formalization results in a set of abstract symbols and well-defined formal rules, which, in an obvious way, have not much to do with the units of natural language.

Inclusion of the metalanguage in the language can be seen as a methodological principle or as an empirical discovery. We will discuss various aspects of this statement by presenting different examples. We are convinced that the principle has deep consequences for linguistics, but that it may take time and research efforts to measure its full impact.

---

1. E.g. Bloomfield (1933).

For Harris, grammar is the formalized description of a given language, say English.

As in any scientific activity, the metalanguage is constructed by the specialists of the field who agree on an object to describe, that is, on facts to be accounted for. Then abstract entities are defined and refined in order to improve the understanding of facts. Consensus among specialists is reached through experiments, but facts and experiments must be *reproducible*. It goes without saying that research programmes should be common to the linguistic community, whether involved in particular language descriptions or in comparing and abstracting descriptions across languages.

Elements of the metalanguage of grammar have been deeply engrained by education among people. Examples are:

– The categories of words such as *verb*, *noun*, *adjective*, *preposition*, *affixes*; more abstract units are the *phrases: noun phrases*, *verb phrases*, etc. and *grammatical functions* such as *subject* or *object*.
– The rules of grammar, such as agreement rules, pronominalization rules, etc.

All of these concepts have been refined into subcategories according to descriptive needs and according to the main application of grammar, which is the teaching of first and second languages.

Most of these concepts are part of a cultural heritage, dating at least to Greek and Roman civilization. Until recently, they have been thought to be universal and have been exported as such by Christian missionaries who used them to describe the languages of Africa, America, Asia, and Oceania. Although specialists have often argued that the Greco-Roman categories are irrelevant to most of these exotic languages, the educational systems of most colonized countries are stuck with this grammatical framework which has been transmitted from generation to generation with remarkable stability.

In fact, the relevance of the Greek-Roman metalanguage even to European languages is far from obvious, but has almost never been questioned. Categories of words have been demonstrated to be useful, for example in the formulation of agreement rules. Confirmation of their value and generality dates back only the nineteenth century, when dictionaries with substantial coverage of the words of a language were built and categories assigned to each word.

1. Sentences

The category *sentence* has a special status as the main object of grammar: a grammar of a language must describe all sentences of a language. Sentences are defined on an intuitive basis. An intuition of grammatical acceptability has been developed independently of meaning and has become the empirical basis of syntactic studies. We recall Chomsky's emblematic example; the string of words:

(1) Colorless green ideas sleep furiously

This has no meaning, but it is grammatical (i.e. it is easy to pronounce, even analyzable in terms of phrases and grammatical functions). The string composed of the same words:

(2) Sleep colorless furiously ideas green

has no meaning either, but in addition no grammatical structure: it cannot be pronounced with a sentence intonation, no grammatical relation can be seen between words. Let us consider another example:

(3) Where fell on the floor?

This is perceived as a sentence (interrogative[2]), but the very similar string:

(4) Where fell on the floor?

is not felt to be a sentence. The string:

(5) the book that fell on the floor

may be recognized as well-formed according to rules we have learned, but is not felt to be a sentence. Experiences shared by linguists have demonstrated the reproducibility of the intuition of sentence acceptability and its limits.
Let us consider examples of the current metalanguage and evaluate them in the framework of grammar construction.

---

2. Another term of the metalanguage.

2. The predicate

The term *predicate* has numerous interpretations by various authors. The introduction of the linguistic notion can be attributed to Aristotle. It was used in a parallel way in logic and grammar in the Middle Ages, it has recently become a technical term in mathematical logic, and has a wide variety of uses in linguistics. There is no use of the term in ordinary language; it clearly belongs to the respective metalanguages of the mentioned fields. The linguistic predicate appears to have the following descriptive use, based on a notion of sentence quite different from the modern one: as mentioned before, sentences are used to provide information by uttering statements about 'things', which can be concrete objects or abstract entities. Hence, a sentence is made of two components: the 'thing' or subject of the statement, and what is said about the 'thing', which is called a predicate. It does not take long to find examples of sentences that are accepted according to the modern definition and for which the analysis in terms of subject and predicate is irrelevant, for example sentences such as *It is six o'clock* or symmetrical sentences such as *Jo is married to Bob* where it is hard to distinguish the role of *Bob* from that of *Jo*. Nonetheless, over centuries, grammarians have attempted to justify in formal terms the two notions: subject and predicate. The notion 'grammatical subject' is a rather operational notion (M. Gross 1999), it has a definition based on agreement rules between the 'thing' of the sentence and the verb, sometimes also called the predicate. Grammarians keep trying to match the notion of grammatical subject with the semantic notion, defining more and more abstract levels of description to meet the ever-growing number of difficulties. For example, since the two sentences *Bob loves Indian literature* and *Indian literature impassions Bob* are more or less synonymous, they should have the same subjects and predicates. In order to arrive at this result, it is necessary to invent an abstract decomposition or representation of both sentences that will satisfy the requirement, something like:

(6) Indian literature causes Bob to be in a state of love

and then assert whatever is desirable: *Bob* is the subject or *Indian literature* is the subject, or both are subjects. Despite centuries of failure, this notion is so firmly established that such analyses have seemed reasonable. Harris demonstrated exceptional intellectual courage in abandoning the notion and adopting for the description of sentences the general schema:

$N_0\ V\ W$

where N_0 is the grammatical subject, V the verb and W the sequence of the complements.[3] This seemingly trivial description has in fact deep empirical consequences:

– It eliminates endless and useless discussions that involve the notion of predicate.
– The representation of sentences is based on the widely-recognized fact that the content of W depends on each verb, and thus, will have to be described case by case.
– An obvious invariant appears for sentences: the sequence $N_0\ V$.

One thus states that all sentences contain a grammatical subject and a verb. The empirical adequacy of this statement has to be discussed. Many languages do not have an obvious agreement rule that adapts a suffix form of the verb to certain changes in the subject; no such agreement phenomenon is observed with complements. Is such a notion of subject appropriate, say, to Chinese, Japanese, or Korean? Also, in Indo-European languages, there are utterances which are clearly recognized as sentences, but which have no grammatical subject or which have no verb. What is their status? Grammarians have given answers in various cases: imperative sentences have a zeroed grammatical subject which can be reconstructed, some impersonal subjects (i.e. *it*) have been 'regularized' under various proposals such as the Extraposition transformation:

(7) *It seems dangerous to act today = To act today seems dangerous*

There are also utterances such as:

(8) *Good night! Merry Christmas!*

(9) *No point going there.*

to which one naturally attaches the notion of sentence but which linguists have often been reluctant to analyze by ellipsis of a verbal unit: *have* in the imperative sentences:

(10) *Have* (*a good night* + *a merry Christmas*).

3. Harris used the symbol Ω for W.

or *there is* in impersonal sentences:

(11) *There is no point going there.*

There are also fully idiomatic utterances, such as:

(12) *The hell with N!*

which are felt to be sentences but resist analyses other than etymological. Such examples are exceptions to the general statement that all sentences have a subject and a verb. But to claim that they are exceptions, two conditions have to be verified:

– It should not be possible to analyze the utterances in question according to the schema $N_0\ V\ W$.
– Their number should be small.

The first condition is not too difficult to check. However, depending on the willingness to adopt (or reject) zeroing rules as a tool for reconstructing regular sources, the outcome of the count of exceptions may change substantially. The second condition is much harder to check. It requires quantitative data of various sorts:

– The number of sentences which do have the regular shape $N_0\ V\ W$ should be known.
– An enumeration procedure for the exceptions should be provided.

Precise quantitative data of both kinds have been obtained for French and for a few other languages. They suggest the construction of syntactic tables for elementary sentences, that is, sentences made of a subject, a verb, and its essential complements, if any.

Approximations are easy to obtain: there are about 15,000 verbs in French that are morphologically simple, that is, made of a single word. In English and in other Romance languages, the number is about the same. This figure takes into account the various meanings of homographic or ambiguous verbs such as *drive* in examples (13–15):

(13) *Bob drives a Ford.*

(14) *What are you driving at?*

(15) *Bob drove Jo to her school.*

Here, we count three verbs *to drive* (there are others). But the figure 15,000

does not include idiomatic or frozen forms such as:

(16) *Bob drove away.*

(17) *The noise drove him crazy.*

which have to be counted separately, since they are composed of two units: *drive* and *away*, *drive* and *crazy*.

In French, we have described more than more than 30,000 frozen sentences, more than twice as many as the number of free sentences. Other important corrections have to be made to these numbers. For example, one has to enumerate N_0 *be Adjective* sentences (e.g. *Jo is tall*), of which there are about 10,000 in French, and others such as N_0 *be Prep* N_1 (e.g. *Jo is in trouble*) for which there is no descriptive tradition (i.e. no name has been given to them); of these there are over 7,000 in French.

Only at this point does the enumeration of utterances that meet the intuitive test of sentencehood yet cannot be analyzed according to the schema N_0 *V W* become a meaningful enterprise. In French, we allowed zeroing operations of the type given above; under these conditions, about 1000 examples of sentences without a subject and/or a verb were found.

Applying the term 'exception' to them is an interesting issue. We have to balance 1000 unanalyzable forms against more than 50,000 regular schemata. Two per cent may appear a reasonable figure for qualifying an event as rare, or it may not. The term exception is in the eye of the linguist, who may vary his point of view. Let us comment on this situation. The sentence forms involving the different verbs *to drive* can be represented by the following schemata:

(18) N_0 *drive* N_1

(19) N_0 *drive* N_1 *to* N_2

(20) N_0 *drive at* N_1

Here, the N_i s are variable noun phrases whose content is semantically constrained by the verb:

– N_0 is **human** in all three cases.
– N_1 is a **vehicle** in (18), is **human** in (19), and is difficult to name in (20).
– N_2 is a **place**.

Schemata can be seen as notational variants for functions of several variables such as:

(21) *drive* (N_0, N_1), *drive* (N_0, N_1, N_2)

At this point, formalization takes on a mathematical character. However, the indices attached to the arguments (i.e. noun phrases) of the verbs are also used to define the syntactic transformations the sentences undergo. For example, passive forms are:

(22) N_1 *be driven by* N_0

(23) N_1 *be driven to* N_2 *by* N_0

For the frozen sentences, the schemata are:

(24) N_0 *drive away*

(25) N_0 *drive* N_1 *crazy*

where the noun phrases N_0 and N_1 are **human** in (24) and (25); N_0 in (25) is either **human agentive** or unrestricted **causative**, that is roughly, sentential. But many examples are such that their frozen parts must be indexed too, because frozen parts may undergo the same transformations as free parts. For example, we have:

(26) N_0 *made up his mind* = *His mind was made up*

It then becomes much less natural and much less convenient to use the formal notation of functions:

(27) *drive away* (N_0)

(28) *drive crazy* (N_0, N_1)

Actually, frozen forms have always been regarded as exceptions. Grammars mention them only briefly and sometimes not at all. However, we have observed a large number of these utterances — as a matter of fact, we observed many more frozen forms than free ones. We are thus entitled to claim that frozenness is a very general phenomenon when it comes to the constitution of sentences. Free variables are not exceptions, since their number is of the same order of magnitude, but they are on a par with frozen items.

Returning to the sentences that do not respect the schema N_0 V W, we have observed that practically all of them were frozen in some sense, hence they are not exceptional from the point of view of sentence formation. Their exceptional character has to be looked for elsewhere, which complicates the situation.

Such a discussion carries us away form the initial question: the interest of

the subject–predicate dichotomy. In fact, from the very beginning, it appears that a cut between the invariant part of the sentence N_0 *V* and its variable part *W* is more meaningful. But in the end, neither cut appears to be significant: the schema N_0 *V W* reflects the general structure of (Indo-European) sentences in a very precise way. Eliminating the notion of predicate leaves us with an improved metalanguage, that is, more operational: the notion of verb is that of a category defined in extension:

(i) Verbs are easy to recognize by their endings and a list of them is easily established.
(ii) Extensions of verbs by prefixation and compounding is productive, and it may be hard to determine which verbs will accept such prefixes as *re-*, *co-*, *un-*, etc. Also, appearance of compounds of the form *to chain-smoke*, *to code-name an operation*, *to radiocarbon-date bones* is not predictable.

Nonetheless, a clear picture of the set of verbs of a language can be reached and used to classify words and processes, leaving a residue that only then can be seen as made of exceptions.

3. Metalinguistic sentences for morphology

Consider the following sentences:

(29) (*The word* + *A word such as*) *arrival* (*takes* + *has* + *contains*) *two* r*'s.*

(30) *Arrival* (*does not have* + *never contains*) *a* y.

From a distributional point of view, subjects contain essentially one variable which ranges over the list of English words. In the same way, the main variable of the complements in (29) and (30) ranges over the English alphabet. The determiner may vary, but within a narrow range, as in:

(31) *English has a capital* e.

Modifying adjectives may be introduced that correspond to some comment made by the utterer of the sentence, as in:

(32) *Oxygen contains a nasty* y.[4]

4. Nasty, because oxide does not.

To be semantically (logically) correct, sentences (29), (31), and (32) must present an identity relation between the spelling of the subject and the letters of the complement. In (30), the relation is different. There are many other analogous sentences that describe the shapes of words:

(33) *Few English words* (*have* + *are of*) *length two.*

(34) *Many English words contain an* e.

(35) *Oxygen begins with an* o *and ends in* n.

Even without introducing the metalinguistic terms 'prefix', 'ending', or 'suffix', such sentences allow a detailed morphological description of written English. By replacing letters with sounds in the preceding examples, the description involves phonemes, that is, it becomes morphophonological. In other words, the sentences constitute the metalanguage of morphology, extensible by the introduction of terms that are more technical. Notice that a device similar to that of morphology is used for expressing **intensity** in:

(36) *He is stupid with a capital* S.

As Harris (1991: 123-144) observed, all these metalinguistic sentences belong to English. At this point, one may argue that they constitute a special subset of English sentences and should not be considered as common English — in other words, that this metalanguage is outside of the language. It seems, however, difficult to sustain such an argument, since many other families of sentences with a similar specialized focus can be easily distinguished. Consider, for example, sentences dealing with **costs** or **prices**:

(37) *This book costs ten dollars.*

(38) *I* (*paid* + *spent*) *ten dollars for this book*, etc.

The direct complements of *to cost* and *to pay* have a highly specialized distribution. The following sentences that correspond to **length measurements** or to **weight** or **time** are similar:

(39) *This river is 100 meters* (*deep* + *long* + *wide*).

(40) *The book weighs five pounds.*

(41) *I spent six hours and 23 minutes reading this book.*

It is hard to say that we are dealing here with technical sublanguages, since

many of these sentences have non-numerical variants that are well-rooted in ordinary language and that are not essentially different, either from a syntactic or a semantic point of view. This is the case for the sentences:

(42) *This river is very* (*deep* + *long* + *wide*).

(43) *The book is heavy.*

(44) *I spent a lot of time reading the book.*

Historically, numerical utterances have become available and have been made precise in a gradual way, following scientific and technical progress. In many cases, they have been allowed to occupy the same syntactic positions as informal utterances also used to express quantities. If we exclude these sentences from the languages, not much will be left to be considered as ordinary, non-technical language. There can be little doubt, then, that the metalanguage of grammar is a part of its subject matter, language itself.

References

Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt & Company.
Gross, Maurice. 1999. "Sur la définition d'auxiliaire du verbe". *Langages* 135: 8–21.
Harris, Zellig S. 1991. *A Theory of Language and Information: A mathematical approach.* Oxford and New York: Clarendon Press.

Chapter 4

On discovery procedures

Francis Y. Lin
St. Hugh's College, Oxford

Zellig S. Harris has often been associated with so-called 'discovery procedures'. This notion refers to procedures which the linguist follows, consciously, to discover the grammar of a language on the basis of linguistic data available to him. It can also refer to procedures which the child uses, unconsciously, to acquire the grammar of a language when it is exposed to the data of the language. So 'discovery procedures' has two senses. Which of these two senses one takes in looking at Harris's formal procedures could play an important role in rejecting or supporting Harris's approach to linguistics.

In this chapter I shall first explore the origin of Harris's procedures and then examine their nature. I shall argue that Harris's procedures were not discovery procedures in the first sense, but that they can be regarded as discovery procedures in the second sense. Chomsky's arguments against Harris's procedural approach will be analyzed in this light: I shall show that those arguments do not really apply to Harris's procedures and that some of the arguments can actually be turned to their favor. Chomsky once also interpreted Harris's discovery procedures as innate procedures for grammar acquisition, but he thought that the procedural approach could not work. His reason will be analyzed in the final section and will be found to be unjustified. The conclusion of this chapter is that we can regard Harris's procedures as innate grammar acquisition procedures and that Harris's approach to linguistics is a valid scientific approach.

1. The Origin of Harris's Procedures

Like most scholarly research, Zellig S. Harris's transformational grammar was developed on the basis of some predecessors' work. A number of linguists before Harris had established a rich tradition of structural linguistics. These

include Saussure, Trubetzkoy, Jespersen, Hjelmslev, Boas, Sapir, and Bloomfield. There is a large literature devoted to the discussion of structural linguistics, e.g. Lepschy (1970), Davis (1973), Hymes & Fought (1981), Newmeyer (1986), and Matthews (1986, 2001). It is perhaps possible to trace the influence on Harris from any of the aforementioned predecessors. But here I shall only discuss Sapir and Bloomfield. After all, it was Harris himself who stated that his work "owes most [. . .] to the work and friendship of Edward Sapir and of Leonard Bloomfield, and particularly to the latter's book *Language*" (Harris 1951a: v), and that "The work here starts off from the distributional (combinatorial) methods of Edward Sapir and of Leonard Bloomfield, to both of whom I am glad to restate my scientific and personal debt" (Harris 1991: vi).

Bloomfield wrote that "It is only within the last century or so that language has been studied in a scientific way, by careful and comprehensive observation" (1933: 3). Before this language studies had been dominated by the desire to fit a language into a certain philosophical scheme. For example, the eighteenth-century scholars "stated the grammatical features of language in philosophical terms and took no account of the structural difference between languages, but obscured it by forcing their descriptions into the schemes of Latin grammar" (Bloomfield 1933: 8). But such studies were misconceived, according to Bloomfield. This is because "Features which we think ought to be universal may be absent from the very next language that becomes accessible" (Bloomfield 1933: 20). What we should do in linguistics, suggested Bloomfield, is to describe individual languages carefully, and then make generalizations on the basis of this description (rather than try to fit languages into certain preconceived schemes). "The only useful generalizations about language are inductive generalizations" (Bloomfield 1933: 20).

On this point Bloomfield was in entire agreement with Sapir. As early as 1921, Sapir had already said that "Each language has its own scheme. Everything depends on the formal demarcations which it recognizes" (1921: 125). And, as he later commented, "The time is long past when grammatical forms and processes can be naively translated by philosophers into metaphysical entities" (1929: 213).[1]

1. See also Jespersen's following remarks:

> Some centuries ago it was the common belief that grammar was but applied logic, and that it would therefore be possible to find out the principles underlying all the various grammars of existing languages . . . Unfortunately, they were too often under the

Both Sapir and Bloomfield made important contributions to the development of modern linguistics, by studying individual languages in a scientific way. What influenced Harris most were their distributional methods. According to Harris, "Sapir's greatest contribution to linguistics, and the feature most characteristic of his linguistic work, was [. . .] the patterning of data" (Harris 1951b: 717). Sapir pointed out that each language has an 'inner' sound system (1921: 57). What is linguistically significant is not what sounds are made in a language but the 'points in the pattern' of the language (Sapir 1921: 58 n. 16). While working with Native Americans, Sapir observed that phonetic differences (however striking) were systematically ignored when they did not correspond to the 'points in the pattern', and were systematically expressed (however subtle) when they hit the 'points in the pattern' of the language of the Indians (1921: 57). Whether a sound is a point in the pattern of a language is determined by all its "specific phonetic relationships (such as parallelism, contrast, combination, imperviousness to combination, and so on) to all other sounds" (Sapir 1925: 48). Sapir used the so-called 'distributional method' to set up linguistic elements (e.g. points in a pattern) on a distributional basis, i.e. the elements are determined relatively to each other. As Harris commented, "The most explicit statement of the relative and patterned character of the phonologic elements is given by Edward Sapir in Sound Patterns in Language" (Harris 1951a: 7 n. 5).

Bloomfield also used the idea of distribution. Take the example of the word *pin* (Bloomfield 1933: 78–79). We can change the first part, in the context of *__in*, and get *fin*, *sin*, *tin*, each partially resembling *pin*. Similarly, we can change the second or the third part of *pin*, and still get partial resemblances. But if we change all the three parts, there will be no resemblance left, as in *pin* and *tack*. This shows that the three parts of *pin* are all phonemes. So, phonemes can be identified in terms of complementary distribution of sounds in certain environments. Similarly, determiners can be identified in terms of the distribution of certain words: "a definite determiner can be preceded by the numerative *all* (as in *all the water*) but an indefinite determiner (as, *some*

---

delusion that Latin grammar was the perfect model of logical consistency, and they therefore laboured to find in every language the distinctions recognized in Latin. (Jespersen 1924: 47)

In the nineteenth century, with the rise of comparative and historical linguistics, and with the wider outlook that came from an increased interest in various exotic languages, the earlier attempts at a philosophical grammar were discountenanced. (*ibid.*)

in *some water*) cannot" (Bloomfield 1933: 203). All English noun expressions are either definite or indefinite, and they can be subcategorized according to the use and non-use of determiners (Bloomfield 1933: 204). English nouns fall into a number of sub-classes. For example, names (proper nouns) occur only in the singular number, take no determiner, and are always definite, e.g. *John*, *Chicago*. And mass nouns never take the article *a* and have no plural, e.g. *the milk*, *milk* (Bloomfield 1933: 205).

Both Sapir and Bloomfield illustrated that there are regularities in a language, such as sound patterns, grammatical processes, form classes, and constructions. Both wanted to investigate such regularities in a scientific way. Both used the distributional method to identify some of the regularities, e.g. points in a pattern of a language (Sapir), phonemes and certain form classes (determiners, nouns, etc.) (Bloomfield). Thus, "Sapir was, with Leonard Bloomfield, a founder of the distributional method which characterizes descriptive linguistics (especially the 'American' school)" (Harris 1968: 766).

But neither Sapir nor Bloomfield sought to define explicit procedures for identifying the regularities in a language. Harris went one step further than they did; he wanted to present a "whole schedule of procedures [. . .] designed to begin with the raw data of speech and end with a statement of grammatical structure" (1951a: 6). He made it clear that descriptive linguistics was not to deal "with the whole of speech activities, but with the regularities in certain features of speech" (1951a: 5). He stated that "These regularities are in the distributional relations among the features of speech in question, i.e. the occurrence of these features relatively to each other within utterances" (1951a: 5). His method of identifying these regularities is the distributional method: "The main research of descriptive linguistics, and the only relation which will be accepted as relevant in the present survey, is the distribution or arrangement within the flow of speech of some parts or features relatively to others" (1951a: 5.).

Harris (1951a, 1970) devised various procedures for identifying the regularities in a language. A typical procedure is the procedure of substitution:

> [W]e take a form *A* in an environment of *C* – *D* and then substitute another form *B* in the place of *A*. If, after such substitution, we still have an expression which occurs in the language concerned, i.e. if not only *CAD* but also *CBD* occurs, we say that *A* and *B* are members of the same substitution-class, or that both *A* and *B* fill the position *C* – *D*, or the like. (Harris 1946: 102)

For example, both *Where did the child go?* and *Where did the young boy go?* are English sentences, so *child* and *young boy* belong to the same substitution-

class. With the procedure of substitution, morpheme classes, sequences of morphemes, intermediate-constituents of phrases and sentences can be identified (Harris 1946, 1963). The idea of 'substitution' was developed later on into the idea of 'co-occurrence', which was used to define and identify transformations (Harris 1957). A fuller explication of Harris's procedures based on substitution or co-occurrence can be found in Lin (2000), which also contains a description of Harris's transformational grammar.

2. The nature of Harris's procedures

Let us now turn to the notion 'discovery procedures'. Harris did propose a set of formal procedures for dealing with the regularities in a language, but he did not speak of them as 'discovery procedures'. (The notion first appeared in Chomsky (1957:51).) What were Harris's procedures for? Were they used to *discover* the regularities (phonemes, morphemes, morpheme classes, constructions, and transformations)? The arrangement of Harris's procedures seems to suggest a positive answer. The procedures first deal with phonology, then with morphology, and then with syntax (though Harris stated that it is not the case that "each procedure should be completed before the next is entered upon" (Harris 1951a:1)). Starting from "the raw data of speech", these procedures yield higher and higher levels of regularities. This process is reminiscent of the practice in physics, where many regularities in the physical world were indeed *discovered*: first there were only certain observable facts, then certain regularities were discovered, then more general regularities were found. This similarity does give one the impression that Harris's procedures were indeed 'discovery procedures', i.e. procedures for the linguist to *discover* the regularities in a language.[2]

Harris did not say explicitly whether his procedures were discovery procedures or not. But in many places in his writings Harris suggested that many regularities had already been known intuitively and that his procedures were a way of replacing those intuitive statements with statements of formal rigor. Harris said that linguists in practice "take unnumbered short cuts and intuitive

2. Chomsky seemed to have capitalized on this similarity, when he argued against the idea of discovery procedures because there are no general procedures of discovering laws in science (e.g. physics). See Section 3 below.

or heuristic guesses", for example, they "will usually know exactly where the boundaries of many morphemes are" (1951a: 1). In this case, Harris's procedures were not intended to discover the boundaries of those morphemes, but the procedures had to yield the same results. "The chief usefulness of the procedures [. . .] is therefore as a reminder in the course of the original research, and as a form for checking or presenting the results" (Harris 1951a: 1).

At another place Harris mentioned a similar point. There are cases where "linguists traditionally use hit-or-miss or intuitive techniques to arrive at a system which works to a first approximation", but this same system can "with greater difficulty — and greater rigor — be arrived at procedurally" (Harris 1951a: 3). These procedures were "cumbersome but explicit", and they were offered "in place of the simpler intuitive practice" (Harris 1951a: 3). The intuitive method is "often based on the criterion of meaning" (Harris 1951a: 8). In some cases a procedure seems to be "more complicated than the usual intuitive method", but "the reason for the more complex procedure is the demand of rigor" (Harris 1951a: 8). Harris (1951a: 186–195, 1954: 780–787) discussed the relationship between meaning and distribution, and he pointed out that many aspects of meaning actually correlate with certain distributional regularities. Thus, in many cases what can be obtained through meaning can also be obtained through formal procedures based on distributional analysis.[3]

Harris's statement that the chief usefulness of his procedures was a form of checking or presenting results suggests that his procedures were not *discovery* procedures, rather they were a way of explaining why those results, which are somehow known independently, are the case. Grammatical constructions and transformations, such as those presented in Harris (1946, 1956, 1957, 1963, 1964, 1965) are known already, even to ordinary speakers. Ordinary speakers seem also to have good knowledge of syntactic categories, i.e. morpheme classes (such as noun, verb, and adjective); they also seem to have good knowledge of morphemes and phonemes (e.g. consonants and vowels). Harris's procedures were not used to *discover* these grammatical entities, because these entities are known independently. But Harris's procedures can be viewed as discovery procedures in another sense. We can assume that the

3. Hiż (1994: 520) comments that "In reality, Harris correlated his formal results with independently known meanings all the time". Nevin (1992: 62) writes: "The aim of [Harris's] methods was not to substitute for these informal ways of coming up with possible analyses, but to verify, for any given result, whether the result had a valid relation to the data of the language".

set of procedures is innately given to the child. So, when the child is exposed in the community of English (or any other human language) speakers, it will use the procedures to discover the grammar, i.e. it will acquire that grammar. With this in mind, let us now examine Chomsky's arguments against Harris's procedural approach to linguistics.

3. Chomsky's rejection of 'discovery procedures'

Chomsky is well known for rejecting 'discovery procedures' and for proposing 'evaluation procedures', and the latter idea then developed into 'universal grammar'. In Lin (2000) I provided a detailed account of the development of Chomsky's linguistic thinking and his relevant arguments. Here I shall summarize some of the main points in that paper and represent them in a logical (though not necessarily chronological) order. Chomsky's arguments against Harris's procedural approach concern issues such as the creativity of language, knowledge of language, the language learning device, the justification of discovery procedures, and the oddity of discovery procedures, etc. Let us see what these arguments are and how forceful they are.

3.1 The 'creativity' argument

Consider the creativity of language first. 'Creativity' means that the speaker of a language can utter a sentence he has never uttered or heard before; he can also judge whether an arbitrary sentence is grammatical or not. 'Creativity' is a fundamental feature of language; any adequate linguistic theory must be able to account for it. But

> structural linguistics has rarely been concerned with the 'creative' aspect of language use, which was a dominant theme in rationalistic linguistic theory. It has, in other words, given little attention to the production and interpretation of new, previously unheard sentences — that is, to the normal use of language (Chomsky 1965: 205, n. 30)

What structural linguistics had been typically concerned with was

> the much narrower problem of constructing several inventories of elements in terms of which utterances can be represented, and had given little attention to the rules that generate utterances with structural descriptions. (Chomsky 1961: 223)

What Chomsky said here was true of structural linguistics prior to Harris. But Harris was interested in the problem of creativity. Harris made it clear that a finite set of kernel sentences and transformations can produce an infinite number of sentences:

> [T]he kernel (including the list of combiners) is finite; all the unbounded possibilities of language are properties of the transformational operations. This is of interest because it is in general impossible to set up a reasonable grammar or description of the language that provides for its being finite. Though the sample of the language out of which the grammar is derived is of course finite, the grammar which is made to generate all the sentences of that sample will be found to generate also many other sentences, and unboundedly many sentences of unbounded length. (Harris 1957: 448)

So Chomsky's 'creativity' argument does not really apply to Harris's transformational grammar.

3.2 The 'knowledge of language' argument

But Chomsky might argue that creativity does not simply mean the ability to generate an infinite number of sentences: creativity should also include the ability to produce and understand sentences that have not been heard or understood before. Take the following typical example, which Chomsky (1986: 8) used:

(1) John ate an apple.

(2) John ate.

(3) John is too stubborn to talk to Bill.

(4) John is too stubborn to talk to.

(5) Whenever the object is missing, an arbitrary object is meant.

Chomsky's argument is this. (1) means that John ate an apple and (2) is understood as meaning that John ate something or other. On the basis of this we may think that the child arrives at generalization (5). (3) means that John is so stubborn that he will not talk to Bill. Now, the object of *talk to* is missing in (4), so (4) is analogous to (2). Thus by analogy, (4) should be interpreted relative to (3) just as (2) is interpreted relative to (1): (4) should mean something like that John is so stubborn that he will not talk to an arbitrary person. But (4) does not mean this, in fact it means that John is so stubborn

that an arbitrary person will not talk to him. Since the child knows how to interpret (1)-(4) without making errors, the just-sketched account in terms of generalization and analogy is therefore not only useless but also wrong.[4]

According to Chomsky, the child understands (1)-(4) correctly, and this shows that it has certain grammatical knowledge. Chomsky argued that this knowledge is innate, and cannot be explained in terms of analogy, generalization, or Harris's procedures of segmentation, substitution, etc. This was Chomsky's 'knowledge of language' argument.

But can't the child's knowledge, shown in this example, be explained another way? Here is another account. There are the following two sentence constructions:

(6) Somebody is too ADJ to do something.

(7) Somebody is too ADJ to do something to.

Sentence (3) is an instance of (6), together with, say, *John is too tired to walk home.* Sentence (4) is an instance of (7), together with, say, *The stick is too hard to break.* The two constructions, (6) and (7), have different meanings. (6) means something like *he is so ADJ that he will not do it*; while (7) means something like *he is so ADJ that people will not do it to him.* (3) and (4) are instances of these different constructions, this explains why they have different meanings.

In this account, it is assumed that child knows the two different constructions and also which sentences are their instances.[5] Adult speakers certainly seem to have this knowledge, so it is reasonable to think that the child somehow acquires this same knowledge in the course of learning English. This knowledge does not need to be innate.

---

4. Chomsky did not put forward this example before 1965, when the war (if there had been one) against Harris's approach had definitely been won. So, this sub-section might seem to be ahistorical. But Chomsky (1965) did use similar examples, such as the difference between *I persuaded John to leave* and *I expected John to leave* (1965:22), and argued that "knowledge of grammatical structure cannot arise by step-by-step inductive operations (segmentation, classification, substitution procedures, filling of slots in frames, association, etc.) of any sort that have yet been developed within linguistics, psychology, or philosophy" (1965:57).

5. Harris-style transformations are needed to account for more complex examples. See Lin (1999, 2000).

3.3 The 'language learning' argument

Chomsky of course would reject the above account of the child's knowledge of grammar. His reason would be this. That account essentially says that the child makes generalizations such as (6) and (7), and establishes an analogy, say, between (3) and *John is too tired to walk home*, and between (4) and *The stick is too hard to break*. That account also implies that the child does not make generalizations such as (5), and does not establish an analogy between (3) and (4). Now, why does the child make certain generalizations or use certain analogies but not others? How does the child know which generalizations and analogies are appropriate and which are not? The fact is that the notions of generalization and analogy, etc. are too vague. This Chomsky sharply pointed out a long time ago:

> [A]lthough there is frequent reference in the literature of linguistics, psychology, and philosophy of language to inductive procedures, methods of abstraction, analogy and analogical synthesis, generalization, and the like, the fundamental inadequacy of these suggestions is obscured only by their unclarity. (Chomsky 1975[1955]:31)

Later on he repeated the same point:

> There is no general notion of 'analogy' that applies to these and other cases. Rather, the term is being used, in an extremely misleading way, to refer to properties of particular subsystems of our knowledge, entirely different properties in different cases. (Chomsky 1988:26–27)

Since no non-empty theory of learning in terms of generalization and analogy exists, the account of knowledge of grammar seems to be hopeless, so Chomsky concluded.

Chomsky's 'language learning' argument is rather forceful. But there is an important distinction to be made between knowledge that something is the case and knowledge of how that knowledge is acquired. For example, I know that the word *table* means the object, but not the size, color or weight. But I do not know how exactly I have come to know that. Not having the latter knowledge does not invalidate the statement that the former knowledge exists. Similarly, we do know that (3) and (4) are instances of (6) and (7), which are different constructions in English. The existence of this knowledge is not affected by the fact that we do not yet know how we acquired it. In Lin (2000) I called this distinction the distinction between 'knowledge-that' and 'knowledge-how'. The implication of this distinction is that knowledge of grammatical constructions and transformations can be investigated separately from the investigation of how they are acquired.

In this light, the account provided in the last subsection is not wrong. We can continue to explain grammatical data in terms of sentence constructions and transformations.[6] It is a separate thing to study how these grammatical entities are acquired by the child. Ideally, these two investigations should go hand in hand, because a grammatical theory would not be very convincing if it ignored the problem of grammar acquisition. How are sentence constructions and transformations acquired by the child? Well, Harris's discovery procedures provide a way — a quite substantial way — of answering this question. We can take Harris's procedures to be psychologically real, i.e. to be innate. We can then try to discover what these procedures are. The procedures Harris proposed are only tentative results. They may be enriched or modified, they may also be replaced by a different set of procedures, which might explain the data better.

3.4 The 'justification' argument

This naturally leads to Chomsky's 'justification' argument. Chomsky thought that the procedural approach tried to find some general, mechanical procedures for discovering the grammar of a language, that is, he thought that those procedures were discovery procedures in the first sense explained at the beginning of this chapter. But as Chomsky pointed out, there were no such general procedures:

> In constructing a grammar for a particular language, one of the decisions to be made concerning each class of sentence is whether to consider them to be kernel or derived sentences. I know of no general mechanical procedure for arriving at the answer to this question, just as I am unacquainted with any *general* mechanical procedure for arriving at a phonemic, morphological, or constituent analysis for the sentences of a language. (Chomsky 1964: 223)

If there are no adequate, general procedures to give us correct results, then what we will need are criteria for developing discovery procedures (in the first

---

6. By 'transformations' (and 'constructions') I here and hereafter mean transformations (and constructions) such as those discussed by Harris. Chomsky also used the term 'transformations', but Chomsky's transformations were very different from Harris's. Harris's transformations were regarded as real entities, not normally falsifiable by further research once established, whereas Chomsky's transformations were theory-internal constructs, which could be modified or even rejected at a later stage of research. See Lin (2000) for more discussion of this issue.

sense of 'discovery procedures') and choosing among them. It was for this reason that Chomsky (1957) proposed 'evaluation procedures' to replace discovery procedures. A major criterion for constructing an evaluation procedure (or measure) for grammars is "that of determining which generalizations about a language are significant ones; an evaluation measure must be selected in such a way as to favor these" (Chomsky 1965:42). This is to say that evaluation procedures must yield "significant generalizations" (Chomsky 1965:45). These generalizations imply that "a person learning a language will select grammars containing these generalizations over other grammars that contain different sorts of generalizations" (Chomsky 1965:45). This thought was later developed into 'universal grammar', two versions of which being the 'principles and parameters' (Chomsky 1981) and the 'minimalist' frameworks (Chomsky 1995). See Lin (2000) for a detailed analysis of this development.

Chomsky pointed out that discovery procedures (in the first sense of 'discovery procedures') need to be justified and evaluated. This was seen to be a weakness of Harris's procedures, and was seen to be a reason for replacing them by evaluation procedures. But in the light of our current discussion (Sections 2 and 3.2 above), Harris's procedures are not discovery procedures in the first sense, i.e. they are not procedures for the linguist to discover the grammar of a language. The grammar of a known language consists of constructions and transformations. These entities are known to the linguist already. Since Harris's procedures are not used for the linguist to discover a grammar, Chomsky's 'justification argument' therefore seems to be off target.

But Harris's procedures can be regarded as discovery procedures in the second sense, i.e. innate grammar acquisition procedures. These innate procedures form the grammar acquisition module. The inputs to the module are grammatical data; and the outputs are constructions and transformations, etc., which are known to us. The task of the innate discovery procedures is to explain why they can yield the desired constructions and transformations when applied to grammatical data. The innate discovery procedures will need to be evaluated, and this is because one set of discovery procedures may explain the acquisition of constructions and transformations better than another set. But this will be normal scientific practice. Chomsky's 'justification argument' will apply in this case, but it will only show that Harris's procedures are scientific if they are seen as innate grammar acquisition procedures. See Section 4 below for more discussion of this point.

3.5 The 'oddity' argument

Let us now consider Chomsky's 'oddity' argument, which says that the idea of discovery procedures is at odds with scientific inquiry. According to Chomsky, a grammar should be treated as "a theory of linguistic intuition" (1965: 19). "There are few areas of science", pointed out Chomsky, "in which one would seriously consider the possibility of developing a general, practical, mechanical method for choosing among several theories, each compatible with available data" (1957: 52–53).

Under the current discussion, grammar is seen to be constituted by constructions and transformations.[7] Most of Harris's constructions and transformations are known already, they are simply formulated from one's knowledge or intuition. So, constructions and transformations, which we know already, are not discovered by using Harris's procedures. Rather, Harris's procedures can be used to explain how we come to know those constructions and transformations. Under the current discussion, grammar is not to be discovered by Harris's procedures, Chomsky's 'oddity' argument against Harris's procedures therefore loses force. The relationship between grammar and Harris's procedures will be further discussed in the next section.

4. Discovery procedures as grammar acquisition procedures

Harris's procedures were not procedures for the linguist to discover the grammar of a known language. But such procedures can be thought of as innate procedures that the child utilizes in acquiring a grammar. This second case is much more interesting, because we really want to understand how the brain works, for example, how the brain enables the child to learn a human language with so much ease. Thus, Harris's procedures can indeed be regarded as discovery procedures — procedures for the child to discover grammar.

It is important to bear in mind the two different senses of 'discovery procedures'. If Harris's procedures are taken to be procedures for the linguist

7. Fillmore also holds that all the constructions in a language constitute the grammar of that language (see Kay & Fillmore 1999). Matthews (1998) argues that the examples Chomsky wanted to explain using universal grammar can be accounted for in a common-sense way (i.e. using constructions and transformations).

to discover the grammar of a language which the linguist already understands or speaks, then Harris's approach will not be feasible. The reason is this. Suppose that a grammar (e.g. the grammar of English) is unknown and that we want to discover what it is. We must then treat the grammar as a scientific theory, whose purpose is to explain grammatical data. It would indeed be odd to devise some mechanical procedures, such as Harris's, for discovering the grammar (theory), because no serious scientists would ever think of such an idea. Notice that grammatical data can be explained by a large number of mutually incompatible grammars (theories). In a more plausible approach one will have to devise criteria for evaluating and choosing among all the possible grammars. Thinking along this line, we will find that it is plausible that the child is endowed with a set of innate grammatical principles, which specify in advance the format of human grammars and which significantly limit the number of possible grammars. Thus, we will find that Chomsky's approach to grammar is much more promising than Harris's. See Lin (2000) for a detailed explanation.

But this situation will be drastically changed if Harris's procedures are taken to be procedures not for the linguist, but for the child, to discover the grammar of a known language (such as English). In this case, the grammar (which consists of constructions and transformations) is already known. What the linguist does is to explain how we know (or how the child comes to know) that there are constructions and transformations in the language. The linguist postulates that there are certain procedures which will, when applied to grammatical data, yield those constructions and transformations. The procedures, but not the grammar, are treated as a theory. The linguist might have to devise different procedures and see which ones work better. Certain procedures which work well at one stage might be revised or even replaced by other procedures at a later stage. The procedures can be thought of as innate procedures in the child's brain, which enable the child to discover the grammar when it is exposed in the language environment. Thinking in this way, Harris's approach is indeed a valid scientific approach, and it should not be rejected.

Interestingly, Chomsky did think of discovery procedures as innate procedures for grammar acquisition. He regarded Harris's procedures of classification, segmentation, substitution, etc. as "an unusually refined, detailed, and sophisticated development of a theory of this general [empiricist] character" (1975[1955]:13), "among the most sophisticated and interesting efforts undertaken within a significant (i.e. nonvacuous) empiricist framework" (1975

[1955]:36), and "an instance of E [empiricism], perhaps the most complex version that has yet been developed" (1975:148). He also suggested that "we [can] interpret the methods of structural linguistics based on segmentation and classification as 'learning theory' (contrary to the intentions of those who developed these methods, so far as I know)" (1975:138; see also 1979:115).

Though Chomsky did regard discovery procedures as grammar learning procedures, he thought that they were "intrinsically incapable of yielding the systems of grammatical knowledge that must be attributed to the speaker of a language" (1965:54), and were "mistaken in principle" (1975:152; see also 1979:116). Chomsky's reason was that discovery procedures "essentially correspond to the empiricist view, according to which the acquisition of knowledge requires operations of classification and induction [and analogy, etc.]" (1979:116), and that such an empiricist approach cannot explain a person's knowledge of grammar. Over the years Chomsky has provided a number of examples to support this claim. In Section 3.2, we saw how he used one of these examples to argue that the empiricist ideas of generalization and analogy are not only useless but also wrong. He rejected the empiricist view that grammar is learned (through procedures of generalization, analogy, and so on) on the ground that such procedures are imprecise and can lead to wrong predictions. He consequently advocated the rationalist idea of innate universal grammar. What he actually achieved was a demonstration that the empiricist learning theory was not yet good enough. But this does not mean that the empiricist view of grammar acquisition should be discarded altogether and replaced by the rationalist view. What is needed is a learning theory which does not merely consist of some vague notions such as generalization and analogy and which can explain the acquisition of a child's grammatical knowledge. As we argued in Sections 3.2 and 3.3 above, grammatical knowledge can be explained in terms of constructions and transformations. Our task now is to find a learning theory which will yield such constructions and transformations given available language data. Harris's discovery procedures have given us a good idea about how to construct such a learning theory.

The empiricist view of grammar learning only makes sense when the grammar of a language is known already. In that case, both the input data (sentences and nonsentences available to a child) and the output (the grammar) are known, and what is left to be found out is the set of procedures which will produce the grammar given the input data. If the output (the grammar) is unknown, then the empiricist view of grammar learning will be

in trouble: we don't even know what is learned, how can we find out the relevant learning mechanisms (procedures)? If the grammar of a language is unknown, then it seems that we will have to take the rationalist approach to grammar, for reasons discussed earlier in this section.

Linguistics has taken a very different route after Harris.[8] Almost all the effort in linguistics has been on working out various grammars (grammatical theories), Chomsky's being only one of them. It is time for the linguist to realize that grammars are known already.[9] What is needed is a theory of how a known grammar, which consists of constructions and transformations, is acquired by the child. Harris's procedures — discovery procedures for the child — are just such a theory. A correct understanding and a real appreciation of Harris's procedures will certainly help us (the linguists) to pay greater attention to discovery procedures than to grammatical theories. For additional linguistic argument supporting this view, and the philosophical rationale behind it, see Lin (1999, 2000).

References

Bloomfield, Leonard. 1933. *Language*. New York: Holt.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague: Mouton & Co.

Chomsky, Noam. 1961. "Some methodological remarks on generative grammar". *Word* 17: 219–239.

Chomsky, Noam. 1964. "A transformational approach to syntax". *The Structure of Language: Readings in the philosophy of language*, ed. By Jerry A. Fodor & Jerrold J. Katz, 211–245. Englewood Cliffs: Prentice-Hall.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.

Chomsky, Noam. 1975[1955]. *The Logical Structure of Linguistic Theory*. New York: Plenum Press. [Typescript described as dating from 1955; the Introduction, 1–53, written in 1975].

8. More precisely, after Harris's 'transformational grammar' period.

9. In this chapter we only analyzed one example of Chomsky's puzzling problems and showed that it can be explain in terms of known constructions and transformations (see Section 3.2 above). Many more such examples need to be similarly analyzed in order to establish the claim that grammars are known already. But the example we used in Section 3.2 is very typical of those in Chomsky's writings (some other examples are discussed in Lin 1999 and 2000). It is therefore reasonable to think that Chomsky's other examples can also be explained away in a similar way. See also Note 7 above.

Chomsky, Noam. 1975. *Reflections on Language*. New York: Pantheon.
Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.
Chomsky, Noam. 1986. *Knowledge of Language: Its nature, origin, and use*. London: Praeger.
Chomsky, Noam. 1988. *Language and Problems of Knowledge: The Managua lectures*. Cambridge, Massachusetts: MIT Press.
Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, Massachusetts: MIT Press
Davis, Philip, W. 1973. *Modern Theories of Language*. Englewood Cliffs: Prentice-Hall,
Harris, Zellig S. 1946. "From morpheme to utterance". Language 22.3:161–183. (Repr. in Harris 1970:100–125, and in Harris 1981:45–70. Page refs. are to the 1970 repr.)
Harris, Zellig S. 1951a. *Methods in Structural Linguistics*. Chicago: University of Chicago Press.
Harris, Zellig S. 1951b. Review of David G. Mandelbaum (ed.), *Selected Writings of Edward Sapir in Language, Culture, and Personality* (Berkeley & Los Angeles: University of California Press, 1949). *Language* 27.3:288–333. Repr. in Harris (1970:712–764).
Harris, Z.S. 1954. "Distributional structure". *Word* 10.2/3:146–162. Repr. in Harris (1970:775–794).
Harris, Zellig S. 1956. "Introduction to transformations". (= Transformations and Discourse Analysis Papers, No.2.) Philadelphia: University of Pennsylvania. Repr. in Harris (1970:383–389).
Harris, Zellig S. 1957. "Co-occurrence and transformation in linguistic structure". *Language* 33.3:283–340. Repr. in Harris (1970:390–457).
Harris, Zellig S. 1963 "Immediate-constituent formulation of English syntax". (=Transformations and Discourse Analysis Papers, No.45.) Philadelphia: University Of Pennsylvania. Repr. in Harris (1970:131–138).
Harris, Zellig S. 1964. "The elementary transformations". (= Transformations and Discourse Analysis Papers, No.54.) Philadelphia: University of Pennsylvania. Excerpted in Harris (1970:482–532).
Harris, Zellig S. 1965. "Transformational theory". *Language* 41.3:363–401. Repr. in Harris (1970:533–577).
Harris, Zellig S. 1968. "Edward Sapir: Contributions to linguistics". In *International Encyclopedia of the Social Sciences* ed. by David L. Sills, vol. 14, pp. 13–14. New York: Macmillan. Uncut in Harris (1970:765–768).
Harris, Zellig S. 1970. *Papers in Structural and Transformational Linguistics*. Dordrecht: D. Reidel.
Harris, Zellig S. 1991. *A Theory of Language and Information: A mathematical approach*, Oxford: Clarendon Press.
Hiż, Henry. 1994. "Zellig S. Harris". *Proceedings of the American Philosophical Society* 138.4:519–527.
Hymes, Dell H. & John E. Fought. 1981. *American Structuralism*. The Hague: Mouton.
Kay, Paul & Charles J. Fillmore. 1999. "Grammatical constructions and linguistic generalizations: The What's X doing Y? construction". *Language* 75.1:1–33.
Lepschy, Giulio C. 1970. *Structural Linguistics*. London: Faber.
Lin, F.Y. 1999. "Chomsky on the 'ordinary language' view of language". *Synthese* 120.2:151–191

Lin, F.Y. 2000. "The transformations of transformations". *Language and Communication* 20.3:197–253.

Matthews, Peter H. 1986. *Grammatical Theory in the United States from Bloomfield to Chomsky*. New York & London: Cambridge University Press.

Matthews, Peter H. 1998. "Should we believe in UG?" *Productivity and Creativity: Studies in general and descriptive linguistics in honor of E.M. Uhlenbeck*, ed. by Mark Janes, 105–113. Berlin: Mouton de Gruyter.

Matthews, Peter H. 2001. *A Short History of Structural Linguistics*. Cambridge: Cambridge University Press.

Nevin, Bruce. 1992. "Zellig S. Harris: An appreciation". *California Linguistic Notes* 23.2:60–64. [Also at http: //informatics.cpmc.columbia.edu/zellig/obit-bn.htm.]

Newmeyer, Frederick J. 1986. *Linguistic Theory in America* (second edition). New York: Academic Press.

Sapir, Edward. 1921. *Language: An introduction to the study of speech*. New York: Harcourt, Brace & World.

Sapir, Edward. 1925. "Sound patterns in language". *Language* 1:37–51.

Sapir, Edward. 1929. "The status of linguistics as a science". *Language* 5:207–214.

Part 2

Discourse and sublanguage analysis

Chapter 5

Grammatical specification of scientific sublanguages

Michael Gottfried
James Madison University

Sublanguages may be considered a consequence of the shared habits of word usage formed in the special activities of various subgroups, part of the 'socio-linguistic division of labor' obtaining in many speech communities.[1] Scientific sublanguages in particular can be characterized as a set of discourses within a particular field of research. In *The Form of Information in Science* (Harris, Gottfried, Ryckman, Mattick, Daladier, Harris & Harris, 1989), the starting point for the investigation was a corpus of sixteen research articles, dating between 1935 and 1970, concerned with the now resolved question of the cellular site of antibody production. The present essay considers some of the complications presented in characterizing the discourses in this corpus and in specifying the sublanguage of immunology. After a summary exposition of the results from that investigation, it proceeds to consider sections of the discourses that are linked to other sublanguages. Other features of these discourses, in particular conjunctions and meta-science operators, are then examined in terms of their relation to sentences describable in terms of the sublanguage grammar. Finally, extensions of the science sublanguage are presented and the hypothesis that the science sublanguage is closed under resolution of referentials is considered in detail.

In the *Form of Information in Science*, discourse analysis and sublanguage analysis were employed to align sentences in articles of the corpus, in a manner which established word classes specific to the subscience, a process

1. Sublanguages as considered here bear an as yet little examined relation to the neighboring concepts of argot, register, style, and dialect in social investigations of language. The types of recurrence-patterns found in various discourses may serve to distinguish text-types or genres, e.g., narratives, scientific observations, questionnaires. (Harris 1982a, p.233)

referred to as 'regularization'. A sample of the word classes and word-class members which are distinguished by these methods follows:[2]

Argument word classes	*Operator word classes*
G antigen, influenzal virus	J inject, administer
A antibody, agglutinin	U stimulate, uptake by
T blood (T_b), spleen (T_s)	V formed by (V_p), appear in (V_i)
C plasma cells (C_z)	W large (W_g), mature (W_m)
B rabbit	

Inasmuch as these classes are constructed on their members occurring in particular sentence-forming operator–argument relations, a set of sentence types is established at the same time. A small sampling of these sentence types follows. Each sentence type on the left is a word-class sequence (with optional word classes in parentheses); in the sample sentences on the right, a bar ('|') separates the corresponding word-class members from one another.

GJ(B)	Toxin | was injected into | the rabbit
AV(C)	Antibody | is concentrated in | lymphocytes
CW(T)	Plasma cells | proliferate in | the spleen

Inasmuch as the grammar is constructed by means of regularizing operations, specific sections of these discourses are closed with respect to these operations. Given a text-sentence such as *Lymphocytes contain antibody*, it may be regularized (aligned with other sentences) as *Antibody is contained in lymphocytes*: both the former sentence and the latter are within the sublanguage.

The word classes and sentence types reported in Harris et al. (1989) provide a grammar for this particular science sublanguage, but not for all of the material in the discourses of the initial corpus. First, there are quantificational operators, e.g., *percent*, *ratio*, and numerals, that serve as local operators on operators of the science-language sentences, as in *The titer of antibody was 8192*. These arithmetic operators and other indicators of quantity cannot be organized as closed word classes in the grammar, and are more appropriately described as belonging to an assumed prior science.

2. Subclasses of these word classes are designated by subscripts. These subclasses are established on the basis of further combinatorial restrictions within the stated classes. Alternatively, they may be specified in definitions given in metalanguage sentences. The metalanguage for a scientific sublanguage is external to the sublanguage.

In addition, many sentences occurring in sections of the articles entitled 'Methods' only contain one of the word classes, and a few sentences in the corpus contain none of the word classes. These sentences generally state various laboratory procedures. As these sentences do not exhibit sentential relations among word-class members, they are not considered part of the sublanguage. However, the immunology sublanguage exhibits various dependencies upon these sentences that merit further study. For instance, the modifier *extracts* on various members of the tissue word class (*T* above) is related to the procedural operator *was extracted*; occurrences of *lymph-plasma* are generally preceded by earlier mentions of centrifugation. One hypothesis to consider is that these procedural sentences can be characterized as belonging to a science sublanguage in its own right with relations to various biological sublanguages.

Aside from those features of the discourses that appear connected to related sublanguages, other features of text-sentences in the 'residue' of the regularizing operations require examination. Sections of the discourse also contain various operators upon the science-language sentences, identified here as *metascience*, which fall outside the established sentence-types. Roughly stated, these sections characterize the investigator's epistemic relation to results reported by science-language sentences. For instance, in *It is possible that* [*plasma cells were contained in the efferent lymph*], the science-language sentence (in brackets) is under the modal operator *it is possible*. Other operators conjoin science-language sentences within the sentences of the corpus. For instance, in *antibody formation was always connected with transitional cells being present in increased numbers*, the conjunctional operator *was always connected with* is identifiable as conjoining an AV sentence *antibody | formation* to a CW sentence *transitional cells | being present in increased numbers*.[3]

The regularized texts are described in terms of sentence-types, e.g., AVC, GJB, CWT, occurring under meta-science operators as well as conjunctional and quantificational operators.[4] The meta-science, conjunctional, and quantificational operators have not been and perhaps cannot be established as

3. Any consideration of the structure of argumentation in science sublanguages requires further examination of meta-science operators and conjunctions. For some preliminary remarks, see chapter 3 in Harris et al. (1989)and Harris (1991) section 10.5.4.

4. More precisely, the regularized texts are described by formulas. Formulas are obtained from sentence-types by specifying designations for subclasses and for local operators upon the operators of the sublanguage.

particular word classes. What is the status of these operators in respect to the immunology sublanguage?

One tack is to include these various operators — meta-science, conjunctional, and quantificational — in the immunology sublanguage. A large corpus of discourses would be needed to establish whether any of these forms could be organized into word classes.

An alternative tack is to restrict the immunology sublanguage to instances of sentence-types ('science-language sentences') and to exclude these operators. In this case, meta-science operators and conjunctions could be considered as a higher-order language. The arguments of these operators would be various pro-forms (usually pro-sentential) referring to the science-language sentences. From a sentence such as *The authors have demonstrated that antibody is in lymphnodes*, one forms the meta-science sentence *The authors demonstrated this* with *this* referential to the science-language sentence *Antibody is in lymphnodes.*

In the example, the science-language sentence *Antibody is in lymphnodes* can be recited alone, i.e., is grammatical. This will not be the case under other operators — contrast the preceding with *The authors have demonstrated* [*titers of antibody in the lymphnodes*], in which the bracketed portion cannot stand alone. Such segments can be established as grammatically independent sentences by further regularization of the texts. For example, the bracketed portion can be denominalized as *titers of antibody are present in the lymphnodes* and can then serve as the referend of a metascience-segment: *The authors have demonstrated this.*

Even if science-language sentences can be established as grammatically self-standing, they often have informational dependencies on the higher operators. That is, under various metascience operators and conjunctions, the science-language sentence is not asserted, but is in various ways negated or stated to be likely, improbable, etc. In *Nothing has emerged which suggests that* [*lymphocytes participate in the formation of antibodies*], the bracketed science-language sentence occurs under a negative metascience operator. In order for formulas of science-language sentences to represent the information contained in the discourse accurately, the assertion status of the science-language sentence needs to be indicated in the formula.

Finally, defining the sublanguage as science-language sentences 'independent of' these meta-operators invites another complication central to the present essay. It presumes either that there are no cross-references from science-language sentences to meta-science material or that such cross-references have been resolved.

The preceding discussion has focused upon ways in which the grammatically characterizable science-language sentences may relate to other features of the discourses in which they occur, features which the sublanguage grammar does not describe. In what ways does the grammar describe *more* than these sentences? The grammar is not a description of a closed text, such as a philologist might produce. The grammar of the immunology sublanguage incorporates a prediction that articles within the sub-field which are added to the corpus will contain sentences analyzable with respect to the established word classes and their combinations.[5] The sublanguage can also be augmented by specifying various closure operations on its sentences — for instance, operators such as *and*, *or*, *not*. Other closure operations may be various rules of inference. Consider the sentence:

> A few scattered plasma cells were found in the retriperitoneal fat in three animals.

A rule of inference which 'drops' modifiers from appositive relative clauses would extend the sublanguage to sentences such as *Plasma cells were found in the retriperitoneal fat*, *Plasma cells were found in the retriperitoneal fat in animals*.[6]

The grammatical description of the immunology sublanguage is largely confined to a sentential analysis of the discourses. A comprehensive description of this scientific sublanguage also requires consideration of the various referential dependencies within and among sentences ('cross-references'). A sample of these cross-references is given below — in the examples, the referential phrase is capitalized and its referend is underlined.[7]

(1) Lymph was separated by centrifugation into lymph-plasma and lymphocytes. THE LYMPHOCYTES contained antibody in higher concentration.

(2) A strain of influenzal virus was injected into the foot-pads of rabbits. After INJECTION, blood was collected from the heart.

---

5. With explicable extensions — see Vieland (1987).

6. Consideration of rules of consequence (see Hiż 1973) is an extension of operator grammar, though one which may be regarded as a natural extension (Harris 1991: chapter 11.4).

7. The passages that follow have been simplified from their occurrence in the original articles in order to focus upon specific instances of cross-reference.

(3) Gross examination showed that the nodes were large and hemorrhagic. THE SAME PICTURE characterized nodes excised on the 5th day.

In the research presented here, the notion of cross-reference is formalized so that resolution of a referential relation can be established and verified. In particular, resolution of a referential yields a sentence that is either a paraphrase or consequence of the text. The notion of cross-reference may be loosely rendered as:

> In a given text, an occurrence of a phrase a cross-refers to an occurrence of a phrase c, with respect to a rule of paraphrase or consequence R, if and only if application of R to the text, with replacement of a by c, yields a paraphrase or consequence of the original text.[8]

From (1), it follows that *The lymphocytes that were separated by centrifugation from lymph contained antibody in higher concentration.* A consequence of (2) is *After a strain of influenzal virus was injected into the foot-pads of rabbits, blood was collected from the heart.* A consequence of (3) is that *Nodes excised on the 5th day were large and hemorrhagic.*

The remainder of this chapter addresses the hypothesis that the science-language sentences are closed under resolution of referentials. If the hypothesis is true, replacement of referentials in these sentences under a rule of consequence or paraphase will yield a consequence or paraphase of the text that is an instance of the science-language types.

Evaluation of this hypothesis requires making two extensions to the science-sublanguage. Firstly, resolution of referentials requires that the sublanguage be augmented by various implicit sentences (in some instances, the implicit sentence may be more appropriately taken as part of some assumed prior science). Consider the referential *the enlargement of the node* in (4):

---

8. This definition, more fully elaborated in Gottfried 1986 (chapter 1) is based upon that presented in Hiż (1969a). Referential relations are here relations between occurrences of phrases within the same text, and as considered here, do not implicate the extra-linguistic relation of 'reference', see Hiż (1969b). Some rules of paraphrase may be formulated in terms of the reduction operations considered in Harris (1982). Frequently, replacements of referends for referentials involve more than a simple substitution operation considered in early transformational analyses, as the referend often requires 'adjustment' to the grammatical category of the referential phrase. See Gottfried (1986: chapters 1 and 3) for further discussion.

(4) There was marked diffuse hyperplasia of lymphoid tissue reaching a maximum two days after the injection. THE ENLARGEMENT OF THE NODE was due to swelling of the cortex.

To obtain a referend for the referential *the enlargement of the node* requires a number of implicit sentences (e.g., *Hyperplasia is of cells. Cells are in a node*), which are part of the assumed science of histology.[9]

Secondly, the set of science-language sentences is augmented by considering some sentences to be incomplete versions of other sentence types. For instance, the sentence *Antigen | was injected* (of sentence type GJ), in which the operator J lacks an argument, is taken to be an incomplete form of another sentence-type, GJB, in which it does not. These incomplete sentences were expanded to include a referential phrase. In the example, the GJ sentence is expanded to include *the animal*, a classifier for various phrases that can serve as the second argument of the J-operator. This results in the GJB sentence *Antigen | was injected in | the animal*. An 'unsaturated' operator such as that in the GJ sentence is termed a 'sublanguage announcer' and the referential phrase that it announces (the B argument in the GJB expansion) is termed a 'zero-referential'. In the example sentence, *injected* is the announcer of the zero-referential *the animal*. In order for science-language sentences to be expanded in this manner, suitable referential phrases need to be established, i.e., referential phrases which can serve as classifiers for members of the argument word classes.[10]

We still face the question whether science-language sentences are closed under resolution of referentials. The 'sublanguage' is here considered to be the 'science-language sentences'. It was hypothesized that resolution of cross-

9. Extension of the sublanguage by implicit sentences permits resolution of referentials in which there is no apparent antecedent. This calls for a qualification of the informal definition of cross-reference presented above so that a referend can occur in a consequence of the preceding text and a set of assumptions, i.e., implicit sentences. See Gottfried (1986: chapter 1, sections 2.4.2 and 5.3).

10. Each of the argument classes in the sublanguage grammar has a classifier. For details, see Gottfried (1986: chapter 3), which demonstrates that the sublanguage could be further regularized in this manner. The present essay does not consider the important question whether classifiers can be established on a distributional basis. James Munz, in Chapter 6 of this volume, illustrates some of the complications involved in answering this question. Related discussion is provided in Daladier, 'Information Units in a French Corpus', chapter 7 in Harris et al. (1989).

references in science-language sentences yields other science-language sentences. This hypothesis was examined by considering the full array of referential relations presented in one article of the corpus, "Influenzal Antibodies in Lymphocytes of Rabbits following the Local Injection of Virus".[11]
In order to evaluate this hypothesis, we first need to consider its presumption that metascience segments and referentials occurring in science language sentences are identifiable as such independently of resolution of cross-references. Referentials as occurrences in science-language sentences are distinguished under three conditions. Firstly, a referential phrase is a referential in a science-language sentence if it is an argument of a sublanguage operator, i.e., a member of the operator word classes J, U, V, etc., or a local operator on members of these word classes. (In the sentence *The production of antibody increased*, *increase* is a 'local operator' on the operator *produced*.) Thus, the many occurrences of, e.g., *the antigen* under *inject*, *the lymphnode* under *present*, *the antibody* under *increase*, etc. are science-language referentials. Similarly, all occurrences of zero-referentials introduced by announcers are science-language referentials.

In some instances, the referential phrase which is the argument of a sublanguage operator is not composed of the recognized vocabulary of the sublanguage. For example, included among science-language referentials are occurrences of the phrases *of the experiment*(*s*)/(*this experiment*) as in:

(5) The greater antibody-titer in lymph and lymphnode-extract than in the serum in the early days of THIS EXPERIMENT has a greater significance.

The occurrence of *the early days* in (5) occupies the same position as *the first 4 days*, *the third day* in similar sentences in which *after* (*the*) *injection* occurs or is announced — thus, *this experiment* is considered a science-language referential.

The second condition under which a referential is a considered to be a referential in a science-language sentence is where the referential phrase is a nominalization of a sublanguage operator. Thus, occurrences of (*the*) *injection* (either explicitly or as reconstructed) are considered science-language referentials, as is the nominalized form *the later rise* in (6):

11. The full citation is S. Harris & T.N. Harris (1949) 'Influenzal Antibodies in Lymphocytes of Rabbits following the Local Injection of Virus', *Journal of Immunology*, 61(2). 193–207. The authors of the article also kindly served as immunology informants.

(6) It may be that THE LATER RISE represents a summation of the declining rate of antibody-production within the node itself plus an increasing rate of antibody from the serum.

Finally, included as science-language referentials are referential phrases which have as their subject an argument word class of the sublanguage. An instance of this situation is presented in (7):

(7) The toxic effect on the local lymphatic tissue was due to THIS PROPERTY of the particular viral agent employed.

Three classes of meta-science segments can be distinguished. Illustrations of each of these cases follow. One class of meta-science operators is distinguished as operators whose second argument is a science-language sentence and whose first (subject) argument is not identical with that of the science-language sentence. An example of this is:

(8) Early investigations indicated that antibodies are present in the regional lymphnode.

In this sentence, the meta-science operator *indicated that* has *early investigations* as its subject and a science-language sentence as its second argument. This group of operators includes: *find*, *study*, *recognize*, *conclude*, *determine*, *see*, etc.

Another group of meta-science operators is distinguished based on the prior determination of subjects for the operators above. These operators have those subjects as their first argument and a member of a sublanguage word class as their second argument, e.g., *were made of* in *Studies were made of the popliteal lymphnode.*

In the final group of metascience operators, the first argument of the operator is a science-language sentence, e.g., *That antibody is released by plasma cells is probable.* Other instances of these operators include: *may be*, *result*, *is possible*, *is significant.*

There are several referential phrases whose status as meta-science segments or as science-language referentials may not be immediately clear, e.g., *the set of observations*, *these quantitative-relations.* Thus, it may be questioned whether *the immunological findings* is determinable as a meta-science or science-language referential independent of its resolution:

(9) THE IMMUNOLOGICAL FINDINGS were correlated with changes in lymphatic tissue.

These referentials are considered parts of meta-science segments. Nearly all of these cases are composed of vocabulary outside of the sublanguage and many are nominalizations of operators such as *find*, *evidence*, *observe*, *demonstrate*, and *conclude*.[12]

Now that it has been established that science-language sentences and metascience-segments can be delimited independently of the resolution of cross-references, we are in a position to evaluate the hypothesis that the immunology sublanguage is closed under resolution of referentials, or, more explicitly, that resolution of referentials in a science-language sentence yields other science-language sentences.

To exemplify some of the details involved, a few passages will be presented. In these passages, referential phrases are capitalized. Referends are established on the basis of semantic judgment. It remains an open question whether or not referends are identifiable on a distributional basis (see Chapter 6 in this volume for some considerations).

(10) Antibodies to influenzal virus appeared in THE LYMPHNODE from two to four days after injection of THE VIRUS into THE FOOT-PAD, whereas normal lymphnodes or lymphnodes derived from rabbits injected with typhoid or dysentery bacilli showed no REACTION WITH INFLUENZAL VIRUS.

In (10) *the lymphnode* refers to a preceding occurrence of that phrase, and *the virus* has as its antecedent *influenzal virus* within that sentence. For each of these referential phrases, one can state a rule of consequence in respect to which replacement of the referential by its referend yields a consequence of the text. This rule states that detaching the last of a string of sentences, one of which contains an anaphoric referential phrase, and substituting the referend for the referential, yields a paraphrase of the text. The referential *the foot-pad* has as its referend the occurrence of *rabbits' feet* in a preceding sentence. Replacement of the referential *the foot* in respect to the stated rule requires an adjustment of *pad* to *pads*. The phrase *reaction with influenzal virus* is a

12. Alternative analyses of these cases are possible. Insofar as these referentials in any case cross-refer to science-language sentences (or sequences of such), their status as meta-science segments or as science-language referentials does not affect the results of the hypothesis. However, these referentials do bear upon the separation of science-language sentences and metascience segments in that they 'separate off' in their referends segments of text-sentences which are metascience from those which are science-language sentences.

referential classifier whose antecedent is *antibodies to influenzal virus appeared.*

(11) Tests with the blood-serum showed no antibody was present *in* THE TISSUE.

In (11), a zero-referential phrase, *the tissue*, is reconstructed. The referend in this case is the complement, *the blood-serum*, of a metascience operator. The example illustrates that referentials in science-language sentences can have as referends occurrences of phrases in metascience segments.[13]
Excerpt (5), repeated below, presents complications.

(5) The greater antibody-titer in lymph and lymphnode-extract than in the serum in the early days of THIS EXPERIMENT has a greater significance.

The occurrence of *this experiment* in (5) poses a potential counterexample to the hypothesis. The immediate referend of this referential is the occurrence of *the experiment summarized in figure 1* in a prior sentence. However, that phrase in turn can be related to an occurrence of a phrase pertaining to an injection.[14]

For the sublanguage of immunology and presumably for other scientific sublanguages as well, the discourses which form the starting point for analysis are seen to be complex structures. Demarcation of this sublanguage requires specification of the way in which the science-language sentences in these discourses are tied to terms relating to procedures as well as to the prior sciences which supply quantificational operators and perhaps the implicit sentences by which cross-references are resolved.

Grammatical specification of the sublanguage requires explicit statement of the operations under which the subset of sentences is closed. The investigation reported here indicates that the science-language sentences are closed under resolution of referentials, substantiating the claim that they comprise an

13. Thus, the hypothesis, entertained in Gottfried (1986), that referentials in science-language sentences do not have as referends occurrences of phrases in metascience segments, cannot be readily sustained. The major 'counter-examples' to this hypothesis consist of referends which occur in complements of the metascience operators above. However, their status as counter-examples to this hypothesis may be questioned. The referends are in turn referential phrases — when the referends of these 'chains' of referential phrases are tracked down, the original referend is found to be housed in a science-language sentence.
14. See Gottfried (1986: chapter 5, section 6.2) for discussion of other complicated cases.

integral structure within the scientific discourses. The cross-referential relations between these two divisions, metascience and science language, are however quite involved and the results suggest that the sublanguage be delimited after these referential relations are resolved.

The relation between metascience and science-language sentences is of especial interest from information-processing and philosophical perspectives. For purposes of information processing, it is critical that the formulaic representations of these sentences are linked to markers indicating whether they are asserted, hypothesized, or in various ways negated. The extension of the sublanguage by implicit sentences also presents issues for information processing of these discourses. There may be ways of generating candidate sentences so that experts in the field can indicate their acceptance.

From a philosophical perspective it is of interest to canvass the considerations involved in extracting metascience operations as distinguished portions of the discourses so that science-language sentences are 'bare' records of the observations. The efforts of members of the Vienna School, most especially of Carnap, were directed at providing for constructed languages a syntactic characterization of observational statements. In the case of the immunology sublanguage the corresponding question is whether one can provide rules of translation that would restrict science-language sentences to reports of observations. These rules would translate a sentence such as *antibodies could be found in plasma cells* to *It was found that antibodies are in plasma cells* in which a meta-science operator is distinguished and the reported fact is stated in the present tense (deletion of the modal *could* requires the support of informants). A related task, also hearkening back to earlier discussions of the language of science, is to clarify the relation that sentences stating procedures have to the sentences of the science language.[15]

15. On the efforts of the Vienna School, see Carnap (1959), and for correctives to misreadings of those efforts, see Friedman (1991). In Gottfried (1986) chapter 5, section 6.3, an attempt is made to 'translate' some paragraphs of an immunology article in this manner. The question of how sentences stating procedures are related to those stating observations is somewhat reminiscent of operationism in the philosophy of science, though, as Hempel (1965) notes, this viewpoint was largely concerned not with statements, but the definition of concepts.

References

Carnap, Rudolf 1959. *The Logical Syntax of Language*. London: Routledge & Kegan Paul.

Friedman, Michael 1991. "The Re-evaluation of Logical Positivism", *Journal of Philosophy* 88:505–519.

Gottfried, Michael. 1986. *Cross-Reference in a Scientific Sublanguage*. Ph. D. dissertation. The University of Pennsylvania.

Harris, Zellig S. 1968. *Mathematical Structures of Language*. New York: John Wiley & Sons.

Harris, Zellig S. 1982a. "Discourse and Sublanguage". In *Sublanguage: Studies of language in restricted semantic domains* ed. by Richard Kittredge & John Lehrberger, 231–236. Berlin: Walter de Gruyter.

Harris, Zellig S. 1982b. *A Grammar of English on Mathematical Principles*. New York: John Wiley & Sons.

Harris, Zellig S. 1988. *Language and Information*. New York: Columbia University Press.

Harris, Zellig S. 1991. *A Theory of Language and Information: a mathematical approach*. New York: Oxford/Clarendon.

Harris, Zellig S., M. Gottfried, Thomas Ryckman, Paul Mattick, Jr., Anne Daladier, T.N. Harris, & S. Harris. 1989. *The Form of Information in Science: Analysis of an immunology sublanguage*. Dordrecht: Kluwer.

Hempel, C.G. 1965. "A Logical Appraisal of Operationism", in *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, 123–133. New York: Free Press.

Hiż, Henry. 1969a. "Referentials", *Semiotica* 2:136–166.

Hiż, Henry. 1969b. "Aletheic Semantic Theory", *Philosophical Forum* 1:438–451.

Hiż, Henry. 1973. "On the Rules of Consequence for a Natural Language" *Monist* 57: 312–327.

Vieland, V.J. 1987. *Can the Language of Science be Formalized?* Ph. D. dissertation. Columbia University.

Chapter 6

Classifiers and reference

James S. Munz
Western Connecticut State University

1. Introduction

The issues surrounding classifier reference have only partly been identified. Classification is a means of establishing common reference of different expressions in the language. Failure to identify classifiers with their classificands often leads to catastrophic failures to identify the coherence of texts. For example, there are articles in pharmacology which deal with small numbers of human subjects. In the body of these articles, individual patients are referred to by their initials but are also referred to collectively as *the patients* or *the subjects.* Failure to recognize the classifier relationships threatens to make sentences using initials irrelevant to those using classifiers and conversely. Distributional analysis of these texts would most likely place both the initials and *patient* and *subject* in the same subclass. If so, the reference route from classifier through classificand would be missed. It is important to develop techniques for identifying or estimating intratextual reference.

Reference is used in at least two different senses. In some cases reference requires denotation. In specifying what a metalanguage is, Harris uses *reference* in this sense.

> Since words can refer to words no less than other things, we can investigate all those sentences which refer to words of the same language. The set of all such sentences in a language is the metalanguage of that language. (Harris 1988:35)

When a phrase (word) is used (as opposed to mentioned), it refers to something, often extralinguistic, by denoting it. Distributional linguistics does not have recourse to extralinguistic denotation. This chapter will take no position on the ontology of referents generally.

In a second sense, pro-words (phrases) are said to refer to other words (phrases) without denoting them. Clearly, Harris does not have this in mind

when he is specifying the sentences of metalanguages. Pronoun reference is reference in this second sense. The problem of pronoun reference has been resolved to a large extent. Part of the meaning of a pro-word for Harris is the zeroed metalinguistic words that denote the words (in the first sense). The same cannot be said for classifier reference in this second sense. Classifiers, as pro-forms, can be taken in the second sense as referring to the linguistic forms they classify whereas in the first sense a classifier refers through its classificand to what the classificand denotes. The usage here will be that a classifier occurrence refers to its classificand and denotes what the classificand denotes. In the same spirit, a reference route in a text goes from classifier occurrences through their classificand occurrences to the denotation.

Relevance is a frankly vague term and will be used here only informally. It has a syntactic sense as when a line in a proof is irrelevant when its removal produces a proof of the same conclusion. More usually it has a non-syntactic sense. Paul Grice's work is of this kind (Grice 1989:27, 31, 86–88). Using conversational postulates he tries to reconstruct for conversational exchanges what would make utterances relevant to their neighbors. This is related to Harris's notion of word-sharing for conjunctional sequences. The issue with classifiers is much simpler. Use of classifiers in texts has on the surface the appearance of changing the topic and creates the appearance of textual incoherence. We will propose introducing hypotheses about classifier relationships which may avoid the appearance of abrupt changes in topic and textual incoherence. This seems to be required for practical machine processing of texts.

Not all occurrences of words which can classify actually do. In connection with homonymity, Harris takes the position that we are dealing with word occurrences (tokens) rather than words (types) (Harris et al. 1987:59). The same is true of pro-words in general and classifiers in particular. In the following, classifiers will be taken as tokens unless otherwise indicated.

Whether or not classifiers are identifiable using distributional methods with sublanguages is unimportant. Classifier relationships (if not distributionally identifiable classifier words) exist in the language generally, and the failure to identify them has the same consequences for textual analysis in any case. This is one of the central points of this chapter.

In Section 3, we develop an example of what appears to be a science sublanguage, one that is not defined by a specific problem and is specified by a wider range of types of literature than that in Harris et al. (1989). Examples of classifiers drawn from texts on cardiac glycosides illustrate some of the convolutions and ephemeral qualities of classifier relations.

The transition to generalized classifiers (those not restricted to sublanguages) is facilitated by a brief discussion of several ersatz modes of reference which linguists normally ignore and which share properties with classifiers. Innovations in classifiers and the ersatz modes of reference are frequent. What is striking is that native readers have so much facility with each type that innovations are usually accepted without even being noticed.

2. Sublanguages and classifiers

Harris contends that classifiers are not identifiable for languages as a whole:

> The criteria for classifiers as operators are that there be no inversion form like (2) [* *A Mammal is a dog.*] and that there be clear dependence of forms like (5) [*A mammal moved in the distance.*] on a source (6) [*Something — said something was a mammal — moved in the distance*]. [. . .]In English these criteria are not satisfied and the status of classifiers is moot. (Harris 1982: 208)

According to Harris classifiers are only recognizable in sublanguages. These criteria are satisfied for at least some sublanguages and, since they are sublanguages, even metalanguages (Harris 1982: 72–73) may have identifiable classifiers.

There is some confusion about what a sublanguage is. Harris first gives what appears to be a sufficient condition: "A subset of the sentences of a language constitutes a sublanguage of that language if it is closed under some operation of the language [. . .]" (Harris 1988: 34). (Notice that he says "if" and not "if and only if".) But he then treats it as if it were also a necessary condition when he excludes the set of elementary sentences as a sublanguage because it does not satisfy the condition (loc. cit). Later he says that sublanguages do not contain their metalanguages (Harris 1988: 38). Still classifiers continue to be vital to textual coherence within a subject-matter area irrespective of whether the closure condition for a sublanguage can be specified over texts in that area. Harris cites the base (unreduced sentences), reduction sentences, and metalanguages as sublanguages (Harris 1988: 37) and says, "Many other subject matters support distinguishable sublanguages." (Harris 1988: 37). The base and reduction subclasses both satisfy the closure condition and do not contain their own metalanguages but are not subject-matter areas. Metalanguages satisfy the conditions as well but have subject-matter areas.

Harris says that the metalanguage is a sublanguage and stipulates that each member of a sequence of higher-order metalanguages is a sublanguage (Harris 1988: 35–36). He does not say more about what constitutes a subject-matter area, or how we identify which ones do and which ones do not support sublanguages. Harris et al. do give an extended example of a science sublanguage (Harris et al. 1989). Several features of the example should give us pause. Harris's example is defined by a specific problem and may not be typical of subject-matter areas generally. The problem is specified by a set of research articles which do not include all the kinds of documents in the area, and some sentences of the articles are excluded (Harris et al. 1989: 10). The distributional methods used to identify subclasses are applied differently, in an important sense, than they are when used to determine the operator grammar for the entire language. The application of distributional methods to science sublanguages doesn't involve interaction with informants. Rather a body of texts is accepted as is. Finally, the techniques employed tend to bury classifiers in the same distributional subclass with what they classify. In such cases, classifiers are not identified but placed in subclasses with their classificands.

Metalanguages excepted, if we start with a subject-area restriction, it is difficult to demonstrate that sublanguage conditions are satisfied. Harris considers an area defined by a body of texts. The area of how antibodies are formed (Harris et al. 1989; Harris 1988: 37–40) is expressed in a collection of research papers. The methodology is different from that used for the language as a whole. The classic method for the whole language involves asking whether a sequence of words is a sentence, and what the acceptability of the sequence is. What qualifies a text for inclusion in the corpus for a sublanguage in Harris et al. (1989) is at least initially a judgment by an approved reviewer or reviewers that it is within the pale of immunology. No questions about likelihood can be asked of an anonymous reviewer. Were we to ask a native reviewer of immunology whether the conjunction of an acceptable sentence with itself is an acceptable sentence, the answer would likely be no. Similarly an unreduced sentence of any complexity is so ugly as to be unacceptable for publication. A sentence with a first person singular pronoun as its subject would not pass review muster.

Showing that the closure condition is satisfied in an interesting sense involves a projection from some corpus. If the projection (that is the determination of what sentences and sequences of sentences to include) is controlled by authors, editors, and reviewers in the subject-matter area, then it is unlikely that the required closure can be demonstrated. Their judgment will be about

whether the sequence is publishable. If the projection is controlled by linguists, then the kind of likelihood gradient that prevents the establishment of subcategories and classifiers will intrude (Harris 1982:72). Harris (1988:38) says, "HCL was washed in polypeptides." will not occur in published laboratory reports, though linguists would give it at least a moderate likelihood. The situation with subject-matter-defined sublanguages is similar to that which confronts linguists when they analyze languages without contemporary speakers. Unless the linguist inserts herself as a speaker, the projection is from published articles and sentences to publishable articles and sentences.

Even if a sublanguage cannot contain its metalanguage, classifiers continue to be vital to textual coherence within a subject-matter area whether or not that area satisfies the condition given above and consequently supports a sublanguage. The subject-matter area developed in the next section may or may not support a sublanguage. I have never been sure how to see if the closure condition is satisfied. It is likely that closure is simply stipulated.

3. Another possible sublanguage

Another plausible candidate for a subject-matter sublanguage that has been investigated is found in studies of cardiac glycosides, a family of compounds related to digitalis. All of the compounds of the class are substitution products of the structure shown in Figure 1.

It is 17H-Cyclo[*a*]-phenanthrene with a furan ring at position 16. Most contain a polysaccharide substituted on the carbon atom labelled *x*. In addition there may be substituents at other positions on the cyclophenanthrene ring structure. It isn't clear whether substituents on the furan ring are allowed. If so there are 12 positions which might have substituents. If not, there are only 9. If there isn't a polysaccharide at *x*, that would be another position for

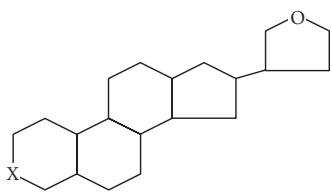


Figure 1. Structure of cardiac glycosides

substituents. The use of *cardiac glycoside* is work in progress. The substituents are hydrocarbons in the known structures, but it isn't clear whether a structure that would be in the family but for a non-hydrocarbon substituent would still qualify as a glycoside. In any case, the class is potentially large.

Rather than being defined by a problem, this subject is defined by a class of compounds. Some but not all members of the class are pharmacologically active. A brief characterization of the scope of the area will be given. It is with this example that we will address some of the problems of classifiers. Initially over eighty articles were selected and over thirty of these were analyzed in detail. The sample covered a wide range of types of article. In addition to research articles of the kind studied by Harris et al., we included review articles and research notes. In many areas of science review articles are common. Some journals are devoted entirely to reviews (e.g. *Pharmacological Review*, *Annual Review of Medicine*). In these extended articles, senior research scientists give summaries of developments in their area. Review articles do not present data or describe research protocols in detail. Introductory sections of research articles are quite similar to review articles, because in setting the research problem they summarize the theories, conjecture, and research of others without necessarily endorsing them. Research notes do not report research protocols or present data in detail but may establish priority or give advice about problems to other researchers.

Virtually the only kind of published literature not included was chapters from textbooks. These appear to be like review articles but much less detailed, and often do not give consideration to minority theories.

The range of topics in the area is quite wide. As with pharmaceuticals generally, there are clinical trials using human subjects and animal studies using in this case cats, dogs, frogs, turtles, squid, pigeons, rats, and guinea pigs, or their parts. There were *in vitro*, *in vivo*, and histological studies. Since most pharmaceuticals are absorbed by a wide variety of tissues, there were studies on papillary muscles, skeletal muscles, smooth muscles, neurons, and erythrocytes. There were studies of drug interactions. There were studies of therapeutic and toxic effects. There was great interest in mechanisms for both therapeutic and toxic effects from the clinical to the cellular level. Since glycosides are absorbed by tissues generally and have similar effects on most tissue types, though the therapeutic effect is exclusively with the circulatory system (treating congestive heart failure), there were a variety of mechanisms studied. So there were articles on effects on the vascular bed, the myocardium, and the vagal system.

At a cellular level there are effects on cation transfer, specifically of sodium, potassium, and calcium. There were also effects on the metabolism of high-energy phosphates.

As might be expected, the boundaries of this subject area are blurred. It is likely that a review of the subject area that Harris et al. studied which was not restricted to central articles would show a similar blurring. One published note was included which was marginal. Glycosides are not uniformly distributed in treated cells. It was important to determine where they attached. This can be done by radiolabeling glycosides and examining treated cells. Histological preparation normally involves dehydration, usually using alcohol. Unfortunately digitalis is alcohol soluble, and so standard preparation would destroy the evidence. The note pointed out that acetone can be used to dehydrate tissue, and digitalis is not acetone soluble. This could have been taken as a note about histology.

The scope of the articles may suggest that the area may not support a sublanguage. This is irrelevant to our investigations of classifiers. Classifiers continue to function referentially in spite of the diversity, and failure to identify them means we lose some of the coherence of the articles.

4. *Ersatz* reference

There are a variety of devices for inter- and intra-textual reference, which are largely taken for granted but have similarities to classifiers. Classifiers may refer to texts or their parts. For example, *articles* may classify and refer to specific bibliographic entities or the articles they name. Readers easily come to terms with unfamiliar reference devices, and further, don't seem to rely on identifiable background information. Similarly, a reader not trained in pharmacology begins to identify classifiers in technical texts after reading a few texts. This suggests that classifiers might in part be identified with their classificands without recourse to background information — a suggestion that will be pursued later.

With regard to intratextual reference, for at least two centuries, until the middle of the nineteenth century, it was common to have marginal descriptions of the contents of sections of the text. These descriptions were set in different fonts than the texts they describe. Distant reference can be done by foot- or end-notes, indicated by superscript numerals or asterisks or sequences of asterisks or other marks. Numerals invoking footnotes may be sequential

for each page, section, chapter, or whole text. If non-numerical superscripts are used, they normally only apply to footnotes for the page. In some cases, because of authors' scruples, style book recommendations, editorial policy, or use of superscripts for other purposes, there may be an unnumbered footnote at the bottom of the first page of an article for such matters as acknowledgments or institutional affiliations of the authors. Such notes are normally in a different font or separated by a line under the text.

Reference to distant parts of a text may be part of the text — e.g., *in chapter 3* — or in parentheses or brackets. Sections may be indicated by a special symbol or by the word "section" or by numerals.

References to other texts have as many modes. The reference may be given in the text. It may be included in a foot- or endnote. If there is an alphabetized bibliography, the reference may be given by citing author and year with a possible letter suffix or by citing author and initials of the title. Either can be set off in commas, parentheses, or brackets. In case there are no numbered footnotes, the bibliography entries may be listed by number in order of first occurrence in the text, and the superscript numerals in the text correspond. The superscript numerals in the text are determined by the numbering in the bibliography. Mattick says, "**N′** includes both names and titles of scientific workers and nouns denoting their work." (Harris et al. 1989: 162). He goes on to say, "This [the device of deriving authors' names from *papers by* ____] allows us to take the word class **N′** as consisting of classifiers for sets of sentences, namely the sentences in (the named) articles." (Harris et al. 1989: 163). The solution is promising but not reconstructable by distributional methods (see below). Further, talk of denotation other than of linguistic material must be taken as not part of the analysis but as marginal comments (reflecting the problem of reference in the first sense noted at the outset).

The point is, with either intra- or extra-textual reference, a variety of devices are available and innovations are easily understood and then become commonplace. Further, none of these devices is subject specific.

5. Classifiers

Digitalis is the grandmother of the cardiac glycosides. It came in to conventional medical use in the eighteenth century when Withering found it being used as a folk medicine. *Digitalis* is the generic name for foxglove. Its chemical structure was not known until well into this century. Digitalis and its offspring

are used to treat congestive heart failure. At the time of the study, at least the following compounds in the family had been found since the original discovery, and by this time many more have: digitalis, digitaline, digitoxin, digitoxigin, monodigitoxin, digitoxoline, lantoside, lantoside C, lantoside E, ouabain, scillaren, scillaren A, hexascillaren A, strophandin, acetylstrophanthidin, g-strophandin and K-strophanthoside.

There is one caveat here; *digitalis* is sometimes used as a generic term for the family of compounds. In the corpus, there is a sequence of over ten articles titled "Studies on Digitalis: [. . .]" most of which do not use digitalis. The following words and phrases are used in one or more of the articles to refer to at least some of the compounds: *agents*, *drugs*, *cardiovascular drugs*, *steroids*, *cardiac glycosides*, *digitalis glycosides*, and *digitalis compounds*. There were other chemical, classificatory relations in the articles. *Ion* classifies *cation* which classifies sodium, potassium, andcalcium, which are referred by *Na*, *K*, *Ca* or *N*a+, *K*+, *Ca*++. *High energy phosphate* refers to adenine nucleotide, neuclotide-H, creatin, CP, ADP, ATP, and APP. *Sympathomemetic amines* refers to epinephrine, l-ephinephrine hydrochloride, nor-epinephrine, l-nor-epinephrine barbitrate. There are three points to be made. First, while many of the chemical classifiers are relatively stable, many are relatively recent. Withering laments the time he spent in the eighteenth century trying to isolate the active ingredient in foxglove tea. Given the state of chemistry at the time and the complexity of the structure, his task was hopeless. Much of the classification was unknown until this century. Second, some of the central scientific classifications are not simple hierarchies but are cross-classifications. For instance, digitalis is both a cardiac glycoside and a drug, but not all cardiac glycosides are drugs (pharmacologically active), and there are certainly other kinds of drugs. Similarly, all glycosides are steroids but not conversely, and some but not all drugs are steroids, constituting another cross-classification. Third, at the margins of most of the scientific classifications there are gray areas where it becomes less and less clear that we are still within the classification. Mattick introduces the notion of background information (Harris et al. 1989:166–167). It is true that many classifiers are part of sciences other than the subscience under study. Obviously the details of such classifications could not be reconstructed from the corpus defining the sublanguage.

Other classifiers are not so well behaved. *Authors* may refer to the authors of the article or authors mentioned in the article. In some cases it appears that a phrase in one traditional grammatical category is referring through a phrase in a different category; so *action* refers to *increase*, *decrease*, *inhibit*. This may

be less a problem in Harris's operator system (Harris 1982:70–103) or even 'optimal transform' of the discourse (Harris 1963:12–13). Metadiscourse phrases frequently refer to what is described in major sections of articles and even entire articles (e.g., *the present results*). In one article *the atrial preparation* referred to an entire protocol for isolating and suspending atria. Another used *paired stimulation* to refer to a complex experimental procedure. The same article uses *this intervention* to refer to paired stimulation. Similarly another used *the two basic preparations* for complicated procedures described in a distant methods section. In one article *variable* refers to oxygen debt; in another *hemodynamic variable* refers to contraction velocity. An animal study used *in all animals* to refer just to the dogs used in the experiment. The following phrases referred through less general phrases: *coronary blood flow*, *this effluent*, *hemodynamic effects*, *effects*, *pharmacological activity*, *abnormalities*, *drug effects*, *toxic effects*, *treated tissue*, *difference*, *abnormal circulatory dynamics*, *these considerations*, *hemodynamic studies*, *this mechanism*, *separate factors*, *this injection*, and *rawolfia alkaloids*.

In addition to some relatively stable and orderly classifiers, there are local situations in which more general phrases are used to refer to what less general phrases refer to. This raises the obvious difficulty that phrases do not wear their generality on their sleeves. These are more likely to be constructed using nontechnical or semitechnical lexical items or phrases. These will be called venerial (hunting) classifiers. They tend to be used for immediate convenience and may be used elsewhere differently. The classifier may appear before or after its classificand. They may not appear in the same sentences and may be used quite some distance apart in the text. Harris recognizes this for the case of pronouns, but notes that the notation used in his example does not indicate extrasentential reference (Harris et al. 1989:18). All the examples used in this section were in different sentences. In the example of the previous paragraph, the dogs were referred to in the methods section of the article at length by *dog*, but were referred to by *animal* only in the results section. It would be easy to multiply examples of both technical and venerial classifiers. The ersatz reference schemata share with classifiers, particularly venerial, that there are many schemata, that new schemata arise frequently, and that they are usually easily recognized by users.

At least half of the articles examined were on cellular level action of the cardiac glycosides and their effects on ion transfer and high-energy phosphate usage. These were certainly more like the set of articles that Harris et al. studied, especially in that there was a protracted disagreement, which was

eventually resolved, over which cations were important in the therapeutic effect. Even in the articles about cellular-level activity both types of classifiers were abundant, and the same is true for most scientific writing. As suggested above, the distinction between technical and venerial classifiers is not sharp. In the other direction, the language defined by the set of English novels analyzed without the use of acceptability gradients, which isn't a sublanguage in any reasonable sense, is replete with classifiers or classifier-like expressions, which are usually easily understood by readers. So the problem of classifiers is not restricted to sublanguages.

Even for sublanguages, where subclasses can be identified, for those cases where the classifier and its classificand are members of the same subclass, there may be no distributional way of determining which way the route of reference runs (because which is more general is not indicated). In this case the classifier has been placed in a class without being identified as a classifier. It is likely that even reference of the classifier (whatever it is) will be missed. Background information will help, but it is striking that readers manage to resolve classifier relationships even when they are not included in background information.

In cases where the classifier is a member of a class and the classificand is a member of a subsubclass of it, the direction of the route to reference can be determined, and the classification relation is distributionally reconstructable, in cases where the classifier is a member of a class and the classificand is a member of a subclass of it, and perhaps only in such cases. Further, in this case the classification relation is distributionally reconstructable.

There are at least two additional difficulties. Consider the case of *drugs* and *cardiac glycosides* as classifiers of digitalis. Digitalis is pharmacologically active and so a drug, but other cardiac glycosides are not drugs — i.e. active. If it is possible to establish subsubclasses for both *drugs* and *cardiac glycosides*, then *digitalis* will be in two subsubclasses. This result is liable to be seen whenever there is nonhierarchial classification. Harris recognizes that in the grammar of the language there will be residual multiple classifications, though they should be avoided if possible (Harris 1982: 68–70). What is quite different here is that this does not necessarily reflect homonymity. Harris does not assume that multiple classification indicates homonymity, though this is an attractive interpretation. Second, in cases like that of *digitalis*, where the term is both generic and specific, it may not be possible by distributional methods to reconstruct the generic reference route because distributional analysis will likely place the word in only one subclass. This is a case of unrecognized homonymity.

There follows a survey of the resolved and unresolved cases of classifier reference in the corpus. Inter-textual reference is not considered. First, consider the case of sublanguages. Classifier types are distributionally identifiable only when the analysis produces more than one level. The methods of producing reference routes are fairly clear as well as their directions. Where only a single level exists, distributional analysis does not identify classifiers. Residues of background classifications — parts of background classifications which appear in the corpus — may be identified as classifiers or not, and behave no differently than other elements of the sublanguage.

Reference to other texts, sections, or (possibly discontinuous) passages remains unresolved not just for classifiers, but generally. Deriving authors' names like *Pfeiffer and Marx* from *papers by Pfeiffer and Marx* (Harris et al. 1989: 163) doesn't solve the problem, though it is a start. Authors' names, bibliographical entries, and titles of papers are not likely to appear in distributional classes. Authors' names and article titles as they appear at the top of an article or in a bibliography (like titles of sections of articles) are not sentential, and would have to be derived from sentences before they could be distributionally analyzed. Mattick's suggestion borders on what was called ersatz reference above. Ersatz references are also involved in reference routes.

Homonymity of classifiers and classificands will be missed if the analysis produces a single-level classification. (The case is the same as for unidentified classifiers.) In case of multiple levels of classes which are not hierarchical (multiple- or cross-classifications), pseudo-homonyms may appear. As used here, a pseudo-homonym is a word which is multiply classified but which is not a homonym. A natural, though not necessarily correct, interpretation of multiple classification is that the two or more similarly spelled words are homonyms. Pseudo-homonyms may lead to spurious reference routes. Missed homonyms may lead to missed reference routes.

For the whole language, Harris asserts that classifiers are not distributionally identifiable. Reference routes exist without classifiers, but without classifiers or classifier-like expressions, there is no way to reconstruct reference routes.

There is a significant area of reference mediated by classifiers or classifier-like expressions which is unresolved. The result of this state of affairs is that a good deal of the coherence of texts will be missed. It is unreasonable to hope that the total, real coherence will be reconstructed, because much of that depends on background knowledge not in the corpus.

6. Extending the distributional analysis

There are possibilities for resolving more of the problem. By enlarging the corpus we may increase the number of levels of subclasses, so that more classifiers (types, not tokens) can be identified, and thence more reference routes and their direction. However, this may also increase the number of pseudo-homonyms.

The second possibility is to go beyond distributional means to enhance 'relevance' connections. Harris allows "limited shortcuts and educated guesses" (Harris 1988: 40) as adjuncts to distributional analysis. What is proposed here would go beyond distributional analysis into what might be pragmatics. The extension is restricted to classifiers (both tokens and types). The proposal is to consider hypotheses of the form either that token *X* classifies token *Y* or that type *X* classifies type *Y*. The case of tokens classifying tokens would be useful in identifying reference routes involving venerial classifiers. The test of an hypothesis would be in the recognition of greater 'relevance' of sentences and 'coherence' of texts and passages. The measure of enhanced coherence is some function of increased number of reference routes. The proposal is not fundamentally different from what is done with pronouns. Of pronouns, Mattick says:

> The derivation makes clear the mechanism of correlation between the two elementary facts, namely the statement of the identity of reference in the two sentences: *Antibodies first appeared in the lymph, which was afferent* ← *Antibodies first appeared in the lymph*; *The lymph was afferent*; *Lymphs are same.* (Harris et al. 1989: 159)

Lymphs are same amounts to an hypothesis about identity of extralinguistic reference. The open problems are (1) constraining the allowed hypotheses, (2) limiting the number of hypotheses allowed, and (3) specifying the function which measures improvement and validating it.

Without restrictions, there are too many candidates for hypotheses. For instance, that this occurrence of *dog* refers to that occurrence of *was* might be a candidate. As the two sets of elements (tokens and types) grow — the set of phrases that can classify and the set of linguistic entities that can be referred to — the exponentially growing number of their combinations becomes unworkable. If the set of possible classificands includes other articles or passages in the current or a different article (Harris et al. 1989: 163), the task is already larger than it appeared. The obvious danger is that the use of unlimited numbers of

unfettered hypotheses might produce a spurious and boring regularity. In the case of pronouns, one linguistic constraint limits the repetition of a phrase. Constraints are necessary, but no work has been done to provide them.

The question of limiting the number of hypotheses allowed is a question of bringing to an end the process of enhancing the distributional analysis. This may be a pseudo-problem. The process is always finite and so must terminate. It is not the goal to turn linguistic analysis into something paralleling literary criticism. Entertaining classificatory hypotheses can be terminated at any point. It might be productive to view the stopping problem as a classic case of Bayesian decision making. There may be more fish in the pond, but you don't have to catch them all.

Perhaps the most daunting problem is developing a measure and giving an argument that it is plausible to take it as a measure of recognized coherence. Since coherence and relevance are presystematic concepts at the moment, a plausibility argument that the measure is acceptable is the most that can be hoped for. The plausibility argument might take the form of proposing functions and seeing in actual cases whether the results of their application agree with presystematic judgments.

References

Grice, Paul. 1989. *Studies in the Way of Words.* Cambridge, Massachusetts: Harvard University Press.

Harris, Zellig S. 1963. *Discourse Analysis Reprints. Papers on Formal Linguistics* 2. The Hague: Mouton.

Harris, Zellig S. 1968. *Mathematical Structures of Language.* New York: Wiley.

Harris, Zellig S. 1982. *A Grammar of English on Mathematical Principles.* New York: Wiley.

Harris, Zellig S. 1988. *Language and Information.* New York: Columbia University Press.

Harris, Zellig S., M. Gottfried, T Ryckman, P. Mattick, A. Daladier, T.N. Harris, & S. Harris. 1989. *The Form of Information in Science: Analysis of an immunology sublanguage*, Boston Studies in Philosophy of Science. Dordrecht: Reidel.

Chapter 7

Some implications of Zellig Harris's discourse analysis

Robert E. Longacre
University of Texas at Arlington

1. Introduction

In this chapter I discuss Zellig Harris's Discourse Analysis in respect to: Discourse Analysis and meaning, the use of transformations in Discourse Analysis, the resultant bidimensional array as a display of the content structure of a discourse, the problem of content transfer in translation, and finally a comparison of Harris's Discourse Analysis with what is currently being done under that rubric, illustrated with a comparison of the analysis of a short text according to both varieties of discourse analysis.

2. Discourse analysis and meaning

Since the publication of Zellig Harris's two initial articles on Discourse Analysis (henceforth DA) in 1952 there has been a discernible shift in his position. He initially emphasized that his DA, like the linguistic analysis of which it was presumably an extension, was distributional and non-semantic in its entirety. Descriptive Linguistics, as he had long propounded, was a study of the mutual distribution of such classes as verbs and nouns, while the DA that he outlined was to be (among other things) a study of the mutual distribution of particular verbs and particular nouns in a given discourse. He emphasized that the methodology was not semantic, i.e., not dependent on the meaning of the items involved in the process, Nor did he affirm that the results of such a DA of a text yielded information as to its meaning: "We may not know WHAT a text is saying, but we can discover HOW it is saying–what are the patterns of occurrence of its chief morphemes." (1952: 1)

He further affirms under section 3.2 Findings: "Various conclusions can be drawn about a particular text by studying the properties of the double array, either directly or in its most simplified form" (1952:26).

But under his section 3.3 Interpretations he skirts around the question as to whether DA yields information as to the *meaning* of the text by insisting that, whatever DA obtains, although it is something more than the bare analysis indicates, questions of text meaning must be assigned to subsequent interpretation:

> The formal findings of this kind of analysis do more than state the distribution of classes, or the structure of intervals, or even the distribution of interval types. They can also reveal peculiarities within the structure relative to the rest of the structure, they can show in what respect certain structures are similar or dissimilar to others. They can lead to a great many statements about the text.
>
> All this, however, is still distinct from An INTERPRETATION of the findings, which must take the meanings of the morphemes into consideration, and ask what the author was about when he produced the text. Such interpretation is obviously quite separate from the formal findings, although it may follow closely in the directions which the formal findings indicate. (Harris 1952:29)

In Harris (1988, 1991) we find meaning handled less diffidently. Thus Harris (1991) contains a major section called Interpretation, with Information a chapter under that head. This chapter begins with the sentence, "Language is clearly and above all a carrier of meaning." (321) While this is certainly a statement which he would hardly have denied in the fifties, it does reflect a certain change of orientation and emphasis. He has a ten-page section (322–332) in this chapter entitled "How words carry meaning", and this section should be of value to teachers trying to assemble a course on semantics. This chapter ends in a next to last sentence which is very quotable:

> And for each word, what the learner or analyzer does is not think deeply about what the word 'really means' but to see how it is used, keeping in mind the fact that the various life-situations in which it used are largely reflected in the various word combinations in which it occurs. [332].

DA itself in the 1988 volume is seen to be a purveyor of information in a quite significant way:

> This book presents a formal method for analyzing the word combinations in articles of a subscience, in a manner that gives the information in the science a more precise form, and may tell a good deal about the structure of the science itself. [first sentence of the Foreword, xv].

Surely, if DA applied to such scientific writing expresses the information of the science in a more precise form and tells us a good deal about the structure of the science itself, we may conclude that the result of DA is insight into the *meaning* of the articles so analyzed.

Actually all this is more of a development on Harris's part than a reversal. Starting out with a repudiation of the use of meaning in DA he ends up with the affirmation of the richness of meaning in the structures which DA uncovers. But this would presumably not be possible if meaning were prematurely introduced into the process itself, even though all along the way the stuff of the analysis is combinations of words as meaningful elements.

3. Use of transformations to standardize a text

Of great significance in all this is Harris's invention of grammatical transformations, his listing of them in detail, to prepare a text for DA. This is evident even in his first published article on DA. The desideratum is to have text lined up in a consistent set of horizontal arrays according to the prevailing N_1VN_2 structure of English. But there is a requirement of identical grammatical structure which is not captured in such a simplistic array; actually, it is required that the first noun be subject and the second some sort of object or complement. I again quote here Harris at length:

> For example, given any English sentence of the form N_1VN_2 (e.g. *The boss fired John*) we can get a sentence with the noun phrases in the reverse order N_2–N_1 (*Jim — the boss*) by changing the suffixes around the verb: *Jim was fired by the boss.* The justification for using such grammatical information in the analysis of a text is that since it is applicable to any N_1VN_2 in English it must also be applicable to any N_1VN_2 in the particular text before us, provided only that this is written in English. (52.4)[1]

1. This really doesn't work across the board for English: *Jim weighs 85 lbs.* cannot be transformed to *85 lbs. are weighed by Jim* simply because N1VN2 runs afoul here of the fact that *85 lbs* in the first sentence is not a regular grammatical object but is a complement indicating measure. In brief, there are functional concerns which limit the application of the formula.

In defense of his method Harris states further down on the same page:

> It merely transforms certain sentences of the text into grammatically equivalent sentences (as N_1VN_2 above was transformed into $N_2V^*N_1$), in such a way that the application of the discourse-analysis method becomes more convenient, or that it becomes possible in particular sections of the text where it was not possible to apply it before.

I note here in passing that DA would be all but impossible without the battery of transformations of which Harris gives here an early version in this chapter but which is expounded more fully in later papers, and which is an integral part of his process in standardizing texts to make them comparable.

In a work intermediate between Harris's beginning articles in the fifties and his volumes published in 1988 and 1991, Harris speaks of 'optimal transforms' (1963:12–13):

> We thus obtain the optimal transform of *D* by selecting particular transformations for each sentence of *D*, and carrying out the equivalences of 2.3 above on the resulting sentences [. . .]. Although the objective is the discourse equivalences, then, the obtaining of the optimal transformation is purely an operation of transformations, and the application of the equivalences is a separate (discourse) operation. The discourse operations can be applied directly to the original *D*, but it will then in general have gaps at points where it might not leave gaps in *D′*.

Nevertheless, as Harris stated in 1952, the bringing in of transformations is clearly a use of information from outside the text under analysis (19). Furthermore, in spite of Harris's rather sanguine assertions to the contrary, DA of the 'raw' untransformed text would be next to impossible. So the DA of a text necessarily depends on used of some information from outside the text. To have admitted this necessary dependence would have imperiled H's initial vision of DA as a purely distributional study within the text.

However, be all this as it may, the fascinating question remains as to what exactly is involved in the all but necessary (I would say, *necessary*) use of transformations in DA? Surely the immediate objective in the use of transformations is to reduce as far as possible the grammatical variety of the text. As much as possible a text is to be reduced to N_1VN_2, or at least to N_1V. Once a text has undergone its optimal transformation, grammar need not be taken further into account; we can now analyze the lexical co-occurrences of the text in their own right. This I take as evidence that in DA of this sort we have not merely extended the grammar to include lexical co-occurrences but have passed from the study of grammar to a third domain (phonology and grammar being the first two domains) into something which could be variously

entitled the content, lexical, or referential structure. The first term is possibly the most neutral. I have used the second term in previous writings (referred to below). Kenneth and Evelyn Pike have developed the third term but along somewhat different lines than in my work.[2]

4. Content analysis: the bidimensional array

Harris's methodology remains the most objective methodology available for getting at the content of a discourse — even granting that DA so conceived is cumbersome and time consuming in its application. It formalizes the study of collocations (or colligations) in a manner not found, e.g., in the work of J. R. Firth and its flowering in the work of Halliday and the others of his school.[3] Moreover, this formalization takes a turn which I find personally congenial as explained below:

From the earliest publications of Harris to his last, one thing is constant in all his work in DA: the result is a double array:

---

2. The Pikes posit as a third mode of discourse what they call reference. They are concerned here with not only the writer and hearer's built-in dictionaries but with their encyclopedias as well. What are the stores of knowledge which are assumed in the text? Exemplary of this approach to a text is Evelyn Pike's 1992 article, "How I Understand a Text — via the Structure of the Happenings and the Telling of Them". The text under analysis is an appeal for funds from an organization called Zero Population Growth. Although this is a hortatory discourse, E. Pike's analysis treats any and all references that are found in the text as a series of 'referential events', developing the referential structure as a kind of implicit narrative. The 39 referential events are each analyzed according to the Pikean paradigm of the four-celled tagmeme with class, slot, role, and cohesion. Then the referential events are organized into vectors each of which is a series of such events. Thus, the references in the text to population-induced pressures become a series of referential events leading to the preparation of a questionaire (the Urban Stress Test) which is formulated and sent out with the results promulgated in a media release, swamping of the staff at ZPG with the response, and a plea for funds to deal with the situation. All these happenings, past, present, and future (e.g., your contribution to the cause) are conceived of as implied in the reader's processing of the data presented in the text. The telling of the events is the grammatical structure of the text as given in which many of the referential events are referred to very summarily, often in noun phrases.

3. I select here Sinclair (1987) "Collocation: a progress report" as representative of an extensive output in research and publication.

> As a product of discourse analysis we obtain a succession of intervals, each containing certain equivalence classes. For a tabular arrangement we write each interval under the preceding one, with the successive members of each class forming a column [. . .]. The very brief text of 2.32 is arranged as follows [I omit some explanatory material in parentheses after certain intervals–REL]:
>
> C S1
> C S2
> C S2
> C S3
> N R0
> N R1
> N R2
>
> The horizontal rows show the equivalence classes present in each interval, arranged according to their order (or some other relation) within the interval. The vertical columns indicate the particular members of each class which occur in the successive intervals.

In the interest of making these symbols meaningful to the reader I note that the text under analysis is one concerning Pablo Casals. The symbols C and S are from the first sentence: *Casals* [. . .] *stopped performing after the Fascist victory.* The suspension points indicate the relative clause *who is self-exiled from Spain* which is transformed into *Casals is self-exiled from Spain* and made a further interval of the *C S* array. Further statements about Casals constitute the following *C S* intervals. The *N R* symbols represent intervals which refer to the recording of Casal's music.

In no way substantially different from such an array is Harris's later work (with six coauthors) — perhaps his crowning achievement — in the 1989 volume *The Form of Information in Science: Analysis of an immunology sublanguage.* In the later work, to be sure, the intervals are more complex internally and there is a greater number of interval types (as seen in the *C S* vs. *N R* intervals above), but the result is nevertheless a bidimensional array obtained by the same methodology — DA proper and resort to a set of transformations to standardize or 'normalize' the texts.

It is not amiss here to point out that Harris's bidimensional arrays in DA are quite parallel to the apparatus used to present a grammatical construction in tagmemics. The following example represents an intransitive clause in Trique (Longacre 1985: 140–141, with slight modification):

$+P_i$:	+S:	$+/-L^2$	$+/-T^2$
Ph_{1i}	$Ph_{11–15}$	$Ph_{11–12}$	$Ph_{21–22T}$
Ph_{23}	$Ph_{21–22n}$	Ph_{41L}	$Ph_{31–33}$
Ph_{32}	$Ph_{81–84}$	Cl_{5L}	Ph_{41T}
$Ph_{4–5}$	M_f		Cl_{5T}
Ph_{6i}	M_{sf}		
	S_{dep}		

Without going into great detail concerning the above array, I immediately point out that it differs in one respect from the bidimensional arrays in Harris's DA: each horizontal row in the latter corresponds to an actual interval in the text, but in the tagmemic bidimensional array, since we are representing not a particular text but all possible constructions of a given type in a language, any item in any column is meant to be combinable with any item(s) in the other column or columns. The intransitive predicate and the subject must be represented in any and all intransitive constructions while locational and temporal expressions need not occur at all, but if they occur, no more than two of each may occur. Some specific rules further restrict permutational and combinatorial possibilities, but these do not concern us here.[4]

Phrases in the 0 decade, i.e. phrases 1–6, are verb phrases; those in the ten decade, i.e. phrases 10–15, are noun phrases. Those in the 20 decade are coordinate and appositional; those in the 30 decade are temporal; those in the 40 decade are prepositional; those in the 80 decade are free pronouns and certain phrasal developments of the same. Subscripts *i*, *n*, *L*, and *T* represent intransitive, nominal, locational, and temporal, respectively. These subscripts are necessary to sort out same structural elements belonging to different

4. Sinclair (1987) has, however, a salutary check at this point. He contends that along with the 'open choice principle' which I illustrate for the Trique intransitive clause above, there exists a second principle, 'the idiom principle' which he defines as follows: "The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear as analyzable into segments." (320). Thus the open choice principle applies to the analysis of "He threw me an orange" while the idiom principle applies to the analysis of "He threw me a curve ball" which, except for a literal meaning in the context of a baseball game, must be construed as a single unit meaning "He referred to me a very difficult/embarrassing problem."

distributional classes. Subscripts *f* and *sf* after *M* (morpheme) indicate fused and semifused subject morphemes whose use is mutually exclusive with free nominal elements; these consist of tone-laryngeal modifications of the last syllable of the verb (in the fused morphemes) plus further possible vocalic or consonantal increments (in the semifused morphemes). S_{dep} indicates dependent sentence.

I believe that the bidimensional arrays used in tagmemics can be considered to represent grammatical syntagmemes, while the bidimensional arrays of Harris's DA represent lexical syntagmemes (G-syntagmemes and L-syntagmemes; see Longacre 1985: 162–166; 1964: 17–19). Both arrays are built on the de Sausurrean relation between syntagmatic (horizontal) and paradigmatic (vertical) relations.

5. The problem of content transference in translation.

In Harris (1988) his chapter 7 concerns translinguistic aspects of his work: "A slightly different form of analysis is employed in Chapter 7 to obtain substantially the same informational units from papers written in French; the French material has not been included in the analysis presented in Chapters 1–6" (xvii). Harris's brief excursion into such translinguistic aspects of his DA work is tantalizing. His work in the linguistic domain of science and in the subscience of immunology and his confining his work to texts concerned with the cellular source of antibodies within that subscience, established that content structure is not markedly different in the texts in two different languages. So far so good. But what if rather than texts in English and French (two contemporary European languages within a common cultural community) and rather than texts expressed in the international idiom of science (wherein precision and consistency of expression are valued) we loosed both of these constraints in our search for common content? What if we were to compare DA of the 'same' document in a European and in a non-European language, and not in the context of science but within that of religious and theological materials — for example, comparing the English or Greek text and a corresponding text in a Mesoamerican Indian language, Trique? Would such an experiment reveal that the search for common content is really illusory in such situations? Would such an exercise reveal that under such conditions — language disparity, the necessity of translating quite freely to achieve intelligibility, in a domain of expression looser than that of science — content

structure evaporates in translation to the point where a version of the Bible is really a quite different book from the original?

As a translator of the New Testament into Trique (published 1966), I prepared the data for such a comparison, and in an article written while the translation was still in progress (Longacre 1958) I essayed such a comparison. That study is based on the almost infinite capacity of language to combine and recombine in context so that the nonidentity of vocabulary grids between any two languages eventually becomes all but irrelevant. In the article I define CONCORDANCE as the occurrence of the same lexical item in various contexts and EQUIVALENCE-CHAIN RELATION as the occurrence of different lexical items in associated contexts according to the specifications of Harris's DA. I then discuss the interplay of concordance and equivalence-chain relation in content structure and how the two relate to each other in the translation process. It was found that while concordance is part of the basis for building equivalence-chain relations, it is essentially less basic to content structure. If a nucleus of the source language (SL) concordance carries over into the translation — in spite of the two vocabulary grids being incommensurate — that is sufficient to expedite the formation of equivalence chains in the target language (TL) document similar to those in the SL document. Loss of concordance does not in itself destroy relations in the TL text if contextual relations according to Harris's methodology relate the disparate items in context. Loss and gain of concordance relative to equivalence-chain relations was discussed in some detail with illustrations involving both Greek and Trique. The results were mildly reassuring to the translator: Of course a translation of the Scriptures is a 'version' in that it inevitably slants and shades the content structure in subtle ways. This is especially seen in local contexts where loss and gain of concordance wipes out small limited equivalence chains or sets up new ones. It is less disturbing in relation to the meaning and thrust of the text as a whole. Certainly, enough of the original text carries through that it is not extravagant to claim that there is a common content to a document regardless of the number and disparate nature of the languages in which it is expressed! That such a startling assertion can be made is largely a byproduct of Harris's DA.

6. Harris's DA and contemporary discourse analysis

Finally something must be said about what Harris termed 'Discourse Analysis' and what others mean by the same term today.

Discourse analysis today includes practitioners whose main concern is intersentential connections, and those with more holistic concerns. Prague School discourse analysis and those continuing their work in theme–rheme analysis (including functional grammar especially as represented by Simon Dik and his followers) represents the first tendency. Those interested in more holistic concerns include those such as myself (Longacre 1996) who operate somewhat as follows:

1. Classifying discourses into *types* such as narrative (a category with very broad range), hortatory (including sermons, pep-talks, advertisements, and appeal letters), expository or descriptive, and others.
2. Recognizing cognitive *templates* on which various discourse types are built.
3. Distinguishing the *backbone* structure of a text of a given type from various sorts of supportive material.
4. Positing climactic or *peak* development which along with initiation and closure of the discourse give it a *profile*.
5. Attempting to account for the various *verb forms* of the language in terms of the above factors.
6. Tracing the distribution of *nouns* and other substantives in terms of participant/thematic roles.
7. Applying a calculus of *intersententional relations* (which have a certain cognitive reality) to relate sentences, or groups of sentences to each other within and across structural paragraphs.
8. Somewhere in this process positing for each text a peculiar *macrostructure* which summarizes the gist or main thrust of the discourse.

I comment in order on each of these factors below but without citing supportive bibliography so as not to swamp the readers with citations:

As to discourse types (1), many of us feel comfortable with an inventory of types such as I suggest. A. Niccacci and others who follow W. Schneider in this and other respects, favor a grand dichotomy into narrative and discourse, using 'discourse' as a general term for all non-narrative text and explanatory/discursive elements in narrative itself.

The recognition of discourse templates or schemata (2) is, at least for narrative, as old as Aristotle. Apart from a bare recital (such as is given by a witness in court) a narrative must have an inciting incident, i.e. some event out of the ordinary, to justify its very existence. Otherwise, the outraged hearer/reader is likely to ask "Why are you telling me this?". Also, it must have plot structure, and that involves a template having at least a climax and a denoue-

ment. Our instinctive feel for plot is evidential of the existence of the narrative template as a cognitive reality. Quite as cognitively necessary is the hortatory template. No exhortation or admonition can be given by one human being to another without some establishment of the right of the exhorter to speak on that particular topic — even if implicitly given in the situation. Furthermore, an exhortation commonly is accompanied by motivation, i.e., warnings, promises, or citation of need. Thus even a rudimentary template with authority of the exhorter, hortatory element proper, and motivation is a part of our cognitive makeup. Other possible templates need not concern us here. I note in passing, however, that not all discourses are built on templates; a notable exception is lyric poetry which is built on a macrostructure but not on a template.

Factor (3) simply recognizes that not all verb forms and clause types are equally relevant to the structure of the discourse. Except in some forms of cryptic and terse poetry we find a structural backbone on which the discourse hangs and which is manifested by a particular verb form or clause type; other verb forms and clause types relate to the backbone as supportive material and represent degrees of departure from the backbone. Thus, in narrative, some sort of perfective or punctiliar tense (or a special narrative tense) marks backbone, and other forms such continuatives, pluperfects, and copulative and descriptive are supportive in various functions.

Factor (4) above, concerning peak and profile in text, is equivalent to saying that any well-formed discourse is not static but proceeds in some direction. Surface structure devices of various sorts can be brought in mark the peak or peaks — and this, in turn, serves to rationalize the use of the latter. To say that a discourse has a profile is to recognize that it is not static whether structurally or semantically.

Factor (5) goes on to claim that all the above factors — text type, backbone versus supportive materials, and placement on a particular point in relation to the template or profile of a discourse — give us what we need to account for the use of various verb types and clause types in the text material in a given language. To cite the greatest distributional contrast, note the occurrence of verb forms which report events (punctiliar/perfective) on the backbone of narrative discourses versus the occurrence of static clause types (i.e., copulative or verbless), or intransitive clauses with inanimate actors on the backbone of expository/descriptive discourses.

Since we must account for the distribution of noun phrases as well as of verb phrases in our texts, some such accounting (6) in terms of participant/theme must characterize discourse analysis as well.

Factor (7) above recognizes that no analysis of discourse that anybody proposes can get by without formulating a set of relations between sentences and groups of sentences in a discourse, whether my own intersentential relations, Grimes's rhetorical predicates, the Beekman-Callow relations posited in their semantic structural analyses, or the extensive set of Thompson-Mann, or even smaller sets such as proposed by van Dijk. Debate continues as to whether such categories are in some sense grammatical or cognitive-semantic, but they are the stuff out of which structural paragraphs are made.

Finally, factor (8), the relevance of a macrostructure to a given text, can be emphasized as a control on what is included in a discourse, what is developed in detail versus what is summarily mentioned, and what is permuted from its natural template order. Actually here Harris's DA — wherein the bidimensional array can be read off to tell us what a discourse is all about — reflects a similar concern.

If Harris's DA gives us the best key yet devised to unlock the content structure of a text, contemporary discourse analysis is largely taken up with the form of the text. But in respect to the positing of templates and intersentential relations as cognitive realities we go beyond the form of the text in insisting that it has in certain respects a correspondence with our cognitive apparatus. This goes beyond merely semantic concerns to what we are and demand as human beings. Furthermore, the structural distinction between backbone and non-backbone is a rough index of the relative semantic import of various strands of the text — although we cannot start with what strikes us as important in a text and what is less important as our point of departure. In most cases, however, the backbone sentences when taken together form a tolerably good abstract of the text. In this way we angle in on the text's meaning as a product of the analysis in a fashion very similar to what happens in a Harris DA.

7. An analysis of a text according to the two varieties of discourse analysis

The text chosen here for the comparison of the two schools of Discourse Analysis is found in the appendix to Harris (1963): "Discourse Analysis of a Story: The Very Proper Gander by James Thurber". This is a compromised narrative by virtue of having a (hortatory) moral at its end, which, however, is not taken account of in Harris's analysis. The story is analyzed by Harris into

GW, HU, and HYG structures, which take off from: GW "[T]here was a very fine gander [. . .] he was strong [. . .] and [he was] smooth, and [he was] beautiful"; HU which takes off from "somebody [. . .] remarked", but all these include remarks about the gander so they are HU (GW); and HYG which tells what other inhabitants of the barnyard did to the gander: "everyone snatched up sticks and stones (against the gander), and descended on the gander's house". Harris assigns the word "propaganda" to a GW interval because "proper gander" is misunderstood or maliciously twisted to "propaganda" and thereby hangs the tale. These three arrays, the GW which describes the gander and his activities, the HU (GW) which tells what people report regarding the sayings of the gander and his purported activities, and the HYG which tells how the other inhabitants of the barnyard drove the gander out of the country, summarize the story quite well. The moral, which Harris did not include in the analysis, doesn't fit too badly: "Anybody who you or your wife thinks is going to overthrow the government by violence must be driven out of the country." In this moral "Anybody" is a G in an extended HU (GW) array and "must be driven out of the country" echoes the HYG — after proper transformations. The analogy is: Just as the gander was driven out, anybody should be driven out if you or your wife (cf. the rooster and the hen) think he's going to overthrow the government by force.

In a discourse analysis of the sort I have summarized above, I would proceed point by point (according to the above eight factors).

(1) The story is a narrative — although compromised by the presence of a moral, a hortatory bit, at the end. A story has no spot on its template for an exhortation, but as as matter of fact, stories in many parts of the world append such a slot as an overt manifestation of a deep structure hortatory intent all through the story. As a matter of fact, many stories have a hortatory disguised structure — since stories may have themes as well as plot — but the skillful writer does not normally need to depend on a moral at the end to point out the theme.

(2) The full template for a narrative discourse is: Stage, Inciting Incident, Mounting tension, Climax, and Denouement. The Stage in this story is sentences 1 and 2: "Not very long ago, there was a very fine gander. He was strong and smooth and beautiful, and he spent most of his time singing to his wife and children".

The Inciting Incident is sentence 3: "One day somebody who saw him strutting up and down in his yard and singing remarked 'There is a very proper gander'." Perhaps the misinterpretation of this remark by the hen as a

reference to "propaganda" in sentence 4 should be part of the Inciting Incident; it at least refers to the Mounting Tension. I want, however, to keep it with sentence 5 so as to relate sentences 4 + 5 and 6 in a dialogue relation. Sentences 7–11, wherein various barnyard fowl, the small brown hen, the duck, and the guinea hen add their derogatory remarks to those of the hen and the rooster in sentences 5 and 6, correspond to Mounting Tension on the template. Sentences 12–13 correspond to Climax; "Finally, everyone snatched up sticks and stones and descended on the gander's house. He was strutting in his front yard, singing to his children and his wife." This is the moment of confrontation, a shift from talking behind the gander's back, who in all innocence suspects nothing coming. Then sentences 14–19 constitute the Denouement, which I will cite below as the peak of the story. In summary, of the narrative template, Stage, Inciting Incident, Mounting Tension, and Climax are present even in a story of this brevity.

According to discourse concern (3) above, we note that in this story, as is customary in English, the backbone of the discourse is carried by clauses whose verbs are in the simple past tense excluding the existential and copulative "was" in the first two sentences. The verb "spent" in the second part of sentence 2 "spent most of his time singing" is evidently a continuative by virtue of the association with the following participle. As to the verbs which propel the story forward, most but not all of them are verbs of speech which introduce quotations: overheard, remarked, told, said, said, remembered, remembered, said, recalled, snatched up, descended on, cried, set upon, and drove (him) out. In sentence 6 the construction "went around [. . .] telling" adds up to a continuative. Several off the line pluperfects occur: in sentence 7: "A small brown hen remembered [. . .] she had seen"; "A duck remembered that the gander had once told [. . .]"; "a guinea hen recalled that she had once seen [. . .]". These pluperfects combine with adverbials expressing distance or indefiniteness to add to the impression of tenuous and somewhat unreliable witness. The small brown hen testifies to having seen the gander "at a great distance". In her second sentence (sentence 8) she vaguely expresses herself by saying "They were up to no good". The duck's witness is that the gander had once told him "he did not believe in anything." Finally, the guinea hen's recollection is that "she had once seen someone who looked very much like the gander throwing something which looked very much like a bomb". Here the pluperfect and the indefinite expressions combine to make the reader sceptical concerning the reliability of the witness of the various barnyard animals. In sentence 13 a past progressive "was strutting", while off the backbone, pictures the innocence of

the unsuspecting gander. In summary, while the past tense forms are the backbone of the story they are supported by clauses with the verb "be", continuatives, pluperfects, and past progressives that flesh the story out.

In respect to factor (4) above, this discourse has a clear peak in sentences 14–19. While sentences 12–13 picture the confrontation: the infuriated crowd descending on the unsuspecting gander and hence may be considered to mark the Climax of the story, sentences 14–19 picture the Denouement. A special surface feature is evident in sentences 15–18; here we not only have quotations without formulas of quotation, we also find that lexical materials in sentences 7, 9–11 are reworked into nominalized compounds used as epithets: *Hawk-lover*, *Unbeliever*, *Flag-hater*, and *Bomb-thrower*.

Factor (5) has been accounted for under (3) above. The verb forms of the language are skillfully used by Thurber in much the way that an artist makes use of his paints. The main features of the design are sketched in the English simple past tense, but existential/copulative *was*, continuatives, and progressives fill in some of the lighter details, while the pluperfect plus expressions of distance and indefiniteness picture some dark details. Finally, the resort to nominalized verbs in compounds in unintroduced quotations marks in flaming red the Climax as a part of the picture which cannot be ignored.

Factor (6) requires our attention to the nouns in the story. The gander is the central participant; he is carefully introduced in the first two sentences, culminating in the anonymous remark in sentence 3: "There is a very proper gander". But, alas, this word of commendation is twisted into one of condemnation as the last two words are interpreted as "Propaganda" in the following two sentences.

In these next two sentences, sentences 4 and 5, first the old hen is thematic, and then the rooster. A small brown hen is thematic in sentences 7–8. A duck is thematic in sentences 9 and 10. Finally, a guinea hen is thematic in 11. There is a variation in the position and form of the formula of quotation in 6 and 9–10, viz., postposed formulas of quotation of VS structure: "said the rooster" in 6, and "said the duck" in 10. In such position and with VS form they close out the quotation and also mark the closing out of the brief domain of the thematic participant involved.

The gander as subject and actor is briefly mentioned in sentence 13, but he becomes object and patient by 19. Such role reversals are common in Climax and Denouement.

The various minor participants and the central participant are each initially introduced with the indefinite article, but take the definite article on

further mention. The rooster is first introduced as her (the hen's) husband on the roost with her, but is on second mention he is referred to as "the rooster"

Pursuant to factor 7 above, I trace out the intersentential relations which are evident in the text; many will turn out to be dialogue relations. I treat any two or more related sentences as, in effect, a (sub)paragraph (for these concerns cf. Longacre 1996: Chapter 4, and for dialogue relations, chapter 5). Some of the relations recognized here are discourse-related, i.e., they relate parts of a discourse as points along the template. Thus the first two sentences constitute the Stage, while sentences 3–11, the body of the discourse, encode the Inciting Incident and Mounting Tension in one surface-structural paragraph. Sentences 12–13 encode the Climax, and sentences 14–19, the Denouement. All other relations are paragraph-level as indicated below.

Sentences 1 and 2 relate to each other by virtue of constituting an *identification* paragraph of which the first sentence is the *thesis* and the second the *identification*. Together they constitute the Stage of the narrative:

> Thesis: Not so very long ago there was a very fine gander.
> Identification: He was strong and smooth and beautiful and spent most of his time singing to his wife and children.

From sentence 3 we are in the body of the discourse as signaled by the words "One day". With sentence 3–11 we encounter an *amplification* paragraph. But this structure takes as its *thesis* another *amplification* paragraph (sentences 3–6) whose *thesis* is sentence 3: "One day someone who saw him strutting up and down in his yard and singing remarked 'There is a very proper gander.'"

The *amplification* of this is a *complex dialogue* between a hen and her husband. Its internal structure consists of an *initiating utterance* (remark) followed by a *resolving utterance* (evaluation) which is reported in a coordinate sentence, the second half of which indirectly reports further verbal activities of the rooster.

The IU (Remark) is a two-sentence unit; sentences 4 and 5 are related, as above, in an embedded *amplification* paragraph whose *thesis* is sentence 4 and whose *amplification* is sentence 5:

> Thesis: An old hen overheard this and told her husband about it that night in the roost.
> Amplification: "They said something about propaganda" she said.

The RU (evaluation) is sentence 6: "'I've always suspected that' said the rooster and he went around the barnyard next day telling everybody that the very

fine gander was a dangerous bird, more likely a hawk in gander's clothing."

The *thesis* expounded by sentences 3–6 is now *amplified* by a *coordinate* paragraph whose *theses* are sentences 7–8, 9–10, and 11. Each *thesis* of this *coordinate* paragraph is expounded by a different speaker — in succession, a small brown hen, a duck, and a guinea hen.

Thesis 1 of this *coordinate* paragraph is again expounded by an *amplification* paragraph composed of sentences 7–8:

> Thesis: A small brown hen remembered a time when at a great distance she had seen the gander talking to some hawks in the forest.
> Amplification: "They were up to no good," she said.

Thesis 2 of this *coordinate* paragraph is expounded by sentences 9 and 11 which constitute an embedded *coordinate* paragraph:

> Thesis 1: A duck remembered that the gander had once told him that he did not believe in anything.
> Thesis 2: "He said to hell with the flag, too," said the duck.

Pulling out of this embedded *coordinate* paragraph and going back to the *coordinate* paragraph whose first *thesis* is sentences 1–8, we recognize a further main *thesis* in sentence 11:

> Thesis 3: A guinea hen recalled that she had once seen somebody who looked very much like the gander throw something that looked very much like a bomb.

Sentences 12–13 give the Climax of the story, introduced by the word "Finally". These sentences constitute an *antithetical* paragraph.

> Thesis: Finally, everyone snatched up sticks and stones and descended on the gander's house.
> Antithesis: He was strutting in his front yard, singing to his children and his wife.

Sentences 14–19 constitute the Denouement in a narrative sequence paragraph of which sentence 14 is the first *sequential thesis* and sentence 19 the last.

The intervening sentences, 15–18, are not necessarily ordered in temporal sequence and may even report simultaneous events; I simply take them together to constitute a *sequential thesis* 2 expounded by a narrative *coordinate* paragraph:

Sequential Thesis 1 (14): "There he is" everybody cried.
Sequential Thesis 2: narrative *coordinate* paragraph
Thesis 1: (15) "Hawk-lover!"
Thesis 2: (16) "Unbeliever!"
Thesis 3: (17) "Flag-hater!"
Thesis 4: (18) "Bomb-thrower!"
Sequential Thesis 3 (19): So they set upon him and drove him out of the country.

According to discourse factor (8) above, we now posit a macrostructure for the story. And indeed this is scarcely difficult: The story is that of a fine gander convicted by hearsay evidence of being subversive and driven out of the country. Thurber's Moral takes the macrostructure, shifts the person references to second person and generalizes: "Moral: Anybody who you or your wife thinks is going to overthrow the government by violence must be driven out of the country."

In concluding this presentation of contemporary discourse analysis I want to mention the factor of *intertextuality*. An obvious instance of this is Thurber's line a "hawk in gander's clothing", which is a reference to Matthew 7:15 "wolves in sheep's clothing". But in a broader sense the very form of the story, the format of an entertaining little animal story, is a form of intertextuality; the moral at the end is specifically reminescent of one of Aesop's Fables. Perhaps the author's approach on adopting this form of narrative is to catch us off guard with something deadly serious? The date of publication of Thurber's *Fables for our Time* (1940) takes us back approximately to the era of McCarthyism when the warning against red-baiting and character assassination would have been quite relevant.

In brief, Harris's DA and contemporary discourse analysis are doing very different things. Harris's DA uncovers and organizes the content structure or the lexicon of the discourse. In this respect it accesses the macrostructure of the discourse in a non-intuitive way. Contemporary discourse analysis, by going at the whole morphological shape of the discourse, highlighting the functions of the nouns and verbs within it, and relating the parts of the discourse to each other, gives us something which could be considered the grammar or syntax of the whole, i.e., how the discourse is put together.

References

Harris, Zellig S. 1952. "Discourse Analysis". *Language* 28.1: 1–30.

Harris, Zellig S. 1963. *Discourse Analysis Reprints.* The Hague: Mouton.

Harris, Zellig S. 1988. *Language and information.* (=Bampton Lectures in America, 28.) New York: Columbia Univ. Press.

Harris, Zellig S. 1991. *A Theory of language and information: A mathematical approach.* Oxford & New York: Clarendon Press.

Harris, Zellig S., M. Gottfried, T. Ryckman, P. Mattick, A. Daladier, T.N. Harris, & S. Harris. 1989. *The Form of Information in Science: Analysis of an immunology sublanguage.* (=Boston Studies in Philosophy of Science). Dordrecht: Reidel.

Longacre, Robert E. 1964. "Prolegomena to lexical structure". *Linguistics* 5: 5–24.

Longacre, Robert E. 1985a. "Items in context: their bearing on translation theory". *Language* 34: 482–491.

Longacre, Robert E. 1985b. "Tagmemics". *Word* 36:137–177.

Longacre, Robert E. 1996. *The grammar of discourse.* 2nd edition. New York: Plenum Press.

Pike, Evelyn G. 1992. "How I understand a text via the structure of happenings and the telling of them". In W. Mann & S. Thompson (Eds) *Discourse Description: Diverse linguistic analyses of a fund raising text* (pp. 227–261) Amsterdam: John Benjamins.

Sinclair, John M. 1987. "Collocation: A progress report". In R. Steele and T. Threadgold (eds.), *Language Topics: Essays in honour of Michael Halliday*, Volume II (pp. 319–331). Amsterdam & Philadelphia: John Benjamins.

Thurber, James. 1940. *Fables for our time.*

Chapter 8

Accounting for subjectivity (point of view)*

Carlota S. Smith
University of Texas

I discuss in this article some of the forms and interpretations usually covered by the phrase 'point of view'. The term is used by linguists for expression of speech and thought, perspective, evidentiality, and other indications of an authorial voice. 'Point of view' is often used almost interchangeably with 'viewpoint,' 'perspective,' and 'subjectivity'. This has led to considerable confusion. In what follows I will be concerned mainly with the notion as discussed by linguists, while recognizing that there is a strong literary tradition.[1] I take it that all expressions of point of view are subjective, since they involve mind; I will use 'subjective,' and 'subjectivity' as general terms rather than 'point of view.'

Subjectivity is expressed by grammatical forms at the sentence level (verbs and their complements, tense, aspectual viewpoint, anaphors, etc.), yet subjectivity arises primarily in discourse contexts. Anaphors and many other forms can only be interpreted with information from outside the sentence. More generally, discourse sets up expectations for structure and interpretation. These expectations depend largely on genre (e.g., narrative, newspaper editorials). The dynamic established by relations between several sentences can set up a pattern that guides interpretation.

One of my goals is to sort out the main types of linguistically conveyed subjectivity as a basis for the systematic interpretation of sentences and discourse. I will distinguish two general classes: sentences that express point of

* I would like to thank the audience at the Conference on Information Structure in Oslo, Norway (December 2000) for discussion of an early version of this material; some of it appears in different form in the Working Papers for the conference.

1. As a literary term 'point of view' refers to presentation of the speech and thought of a fictional character or, more generally, as "the perceptual or conceptual position in terms of which narrated situations and events are presented" (Prince 1987:73).

view, and perspectival sentences that present a situation from a particular standpoint. Both involve the mind and are therefore subjective, but in clearly different ways. The discussion is mainly about English, though I also comment on other languages.

I also consider the question of how grammatical forms give rise to interpretations of subjectivity when they occur in sentences. I suggest a 'composite' account: subjectivity is conveyed by composites of syntactic and semantic factors and interpreted by rules which look at several grammatical forms together. Typically the forms convey more than one kind of information. The composite approach is well-suited to point of view because it deals naturally with such variety. I will suggest principles for interpreting subjectivity in the framework of Discourse Representation Theory.

Section 1 sets the stage with introductory examples and commentary; Section 2 organizes the phenomena and discusses the basic distinctions; Section 3 sketches the composite analysis; Section 4 concludes.

1. Introduction

The examples below illustrate different types of subjectivity. The first three examples are in the first person. Fragments (1a–b) are from novels; (1c) is from a newspaper article of opinion.[2]

(1) a. . . . 1 "My God," Alec said. "What is he doing?" 2 "Who?" 3 "Your boss," Alec said. "Standing in the fountain." 4 I crossed to the window and stared downwards: down two floors to the ornamental fountain in the forecourt of the Paul Ekaterin merchant bank. 5 Down to where three entwining plumes of water rose gracefully into the air and fell in a glittering circular curtain. 6 To where, in the bowl, calf-deep, stood Gordon in his navy pin-striped suit . . . (Dick Francis)

b. . . . 1 I sipped my drink and nodded. 2 The pulse in his lean grey throat throbbed visibly and yet so slowly that it was hardly a pulse at all. 3 An old man two-thirds dead and still determined to believe he could take it. (Raymond Chandler)

c. . . . I feel reasonably certain of the final verdict on the current impeachment affair because I think history will see it as the climax of

2. Text sources are listed at the beginning of the references, below.

a six-year period marred by a troubling and deepening failure of the Republican party to play within the established constitutional rules. (Peter Ehrenhalt)

I note briefly the forms that indicate subjectivity in these fragments. (1a) begins with direct quotations, introduced by the verb *said* and quotation marks. The perception verb *stared* appears in sentence 4; we understand what follows as expressing what the subject perceives — including the material in sentences 5 and 6, within the scope of *stared*.[3] Perception is involved in (1b), but less directly: sentences 2 and 3 suggest it although no perception verb appears. The fragment in (1c), with the verbs *feel* and *think*, expresses the speaker's feelings and thoughts. In all three we would ascribe a point of view to the referent of the first person pronoun — the speaker-reporter.

The next example is from a detective novel in the third person. Consider the interpretation of sentence 4, which is not preceded by a verb of thought or perception.

(2) 1 The workshop was dusty and messy, crusted with bits of old clay, and it suited a part of Mara's personality. 2 Mostly she preferred cleanliness and tidiness, but there was something special, she found, about creating beautiful objects in a chaotic environment. 3 She put on her apron, took a lump of clay from the bin and weighed off enough for a small vase. 4 The clay was wet. (Peter Robinson)

Sentence 4 may express a perception or awareness of Mara's, the person working with the clay; or, it might be due to the reporter. This kind of uncertainty as to who is responsible for a phrase or clause is not uncommon. The next example illustrates subjectivity of a different kind: an authorial voice that evaluates and suggests a temporal standpoint. (2) is a from an article of popular science:

(3) *Cell Communication: The inside story*
1 It might seem surprising that mere molecules inside our cells constantly enact their own version of telephone without distorting the relayed information in the least. 2 Actually, no one could survive without such precise signaling in cells. 3 The body functions properly only because cells communicate with one another constantly. 4 Cells

3. Following Harris, one might account for this inclusion by the presence of zero allomorphs of *stared* etc. in each sentence.

> of the nervous system rapidly fire messages to and from the brain. 5 But how do circuits within cells achieve this high-fidelity transmission? 6 For a long time, biologists had only rudimentary explanations. 7 In the past 15 years, though, they have made great progress in unlocking the code that cells use for their internal communications. (Scientific American)

The sense of voice can be traced to a number of sources: the modal *might*, the verb *seem*, the predicate *surprising*, all in sentence 1; the adverb *actually* in sentence 2; the direct question which comprises sentence 5; the time adverb *in the past 15 years* and the concessive adverbial *though* of sentence 7.

The well-known examples of (4) have a subjective element, suggesting a particular standpoint from which the situation is presented:

(4) a. Physicists like yourself are a godsend.
b. John pulled the blanket over himself.

(4a) takes the perspective of a speaker toward an addressee (from Ross 1970); (4b), from Kuno (1987), demonstrates what he calls an "empathy perspective", in which the reflexive indicates the perspective of John, a participant in the situation.

These examples show that a variety of linguistic forms convey subjectivity. This diversity is essential to the composite analysis sketched in section 3. When a sentence expresses a point of view, or takes a particular perspective, we need to identify the person/mind responsible. The question is, to which mind do we ascribe the material expressed or implied? Ascribing responsibility is part of the interpretation of a sentence that conveys subjectivity.[4] In the sentence *Mary believed that John was sick*, for instance, we ascribe the belief expressed in the complement clause to Mary and not to the reporter, or to John (assuming that we trust the reporter). In the sentence *Mary unfortunately may win the race* we ascribe responsibility for the modal and the adverbial commentary to the reporter. We have also seen cases where either the reporter or the participant in the situation may be responsible. For 'neutral' sentences — by convention objective — the question of responsibility does not arise. We

4. In narrative fiction, it is customary to distinguish sentences that contain a 'self' at a moment corresponding to an act of consciousness (Banfield 1982: 158). Sentences without a 'self' report events "objectively, independent of an explicitly-perceived, narrating self" (Fleischman 1991: 31).

may ask whether a neutral sentence is true but not, usually, who is responsible for it.

2. Subjectivity: point of view and perspective

In this section I discuss different types of subjectivity and the key linguistic factors that convey them. I suggest two major categories: sentences expressing a point of view, and sentences that take a particular perspective on a situation. Among sentences that express point of view, I distinguish three subclasses. They have been developed on the basis of intuitions about meanings and the linguistic forms that appear in them. The subclasses of point of view are communication, contents of mind, and evaluation. Perspectival sentences fall into two subclasses: sentences that involve perception and those that do not. The focus of this discussion is grammatical: I do not consider lexical choice, although that is clearly an important factor. The particular words that appear in a sentence may strongly suggest choices that a particular participant would make, and thus provide access to the mind of that participant.

2.1 *Expressions of point of view*

In this section I consider the expression of point of view. Since the forms are sententially based, the discussion is in terms of sentences; however, factors in the linguistic context are often relevant. This will become clear directly. I begin with point of view as expressed in sentences involving communication, and then move on to sentences expressing content of mind and evaluation.

2.1.1 *Communication*

This category of expression deals with speech and other communicative events, which are by definition external. Linguistic presentations may consist of quoted speech, represented speech, or indirect speech. The first two directly present what was said, often introduced by a verb of communication. Quoted speech is often referred to as 'direct' speech. I use the term 'quoted' here because I want to recognize quoted and represented speech as directly presented communication. In contrast, indirect speech may involve the reporter's recoding of the communication.

The features of communicative sentences have been analyzed extensively (Partee 1975; Banfield 1982). Verbs of communication are a distinct syntactic

class. They allow a direct object complement which expresses the actual communication and an indirect object referring to the addressee (X said Y to Z). The class includes the verbs *say*, *ask*, *request*, *command*, *declare*, *confess*, *advise*, *insist*, *claim*, *shout*, *read*, *sing*, *remark*, *observe*, *note*, *yell*, *swear*, *promise*, *announce*, *pray*. Some verbs of thought have the same syntactic characteristics. I begin with examples of quoted and indirect speech. The former reproduces what was said, while the latter is indirect, a report. In discussing these and all later examples I will refer to the source of a sentence or main clause — the writer or speaker, the person responsible for the quotation or report — as the reporter.

(5) Quoted speech and thought
   a. "I am getting ready for the party this afternoon."
   b. Mary_i said "I_i am excited."
   c. Mary_i told me_j yesterday at the station, "I_i will meet you_j here."
   d. Mary_i asked "Do I_i have to go?"

(6) Indirect speech and thought
   a. Mary_i told me_j yesterday at the station that she_i would meet me_j there.
   b. Mary_i asked whether she_i had to go.

Quoted speech is just that: as in (5), it typically appears with a verb of communication and a direct representation of what was said. Person, tense, and other deictics orient to the first-person speaker, as in the original utterance. In contrast, indirect speech does not present exactly what was said: pronouns, tense, and deictics shift in accord with the report of the communication. Compare for instance (5c) and (6a), from Banfield (1982). The complement of (5c) reproduces Mary's utterance: it has the first person, present tense, and a proximal deictic oriented to the speaker. In (6a) these forms are shifted: the tense is past in accord with the tense of the main clause verb, the pronoun is third person, the deictic pronoun is non-proximal (distal). Syntax also shifts in indirect speech: constructions that are limited to main clauses, such as questions and exclamations, have other forms (or do not appear at all). Compare for instance examples (5d and 6b): the question is direct in (5d) and indirect in its counterpart (6b).

The tenses of main and complement verbs are in concord, as the examples of (6) show; this kind of concord is also known as sequence of tense (Comrie 1986). It is required in English, but not in all languages. Japanese and Navajo,

for instance, do not have shifted deictics nor tense concord in the complements of communication verbs. (Hirose 2000, Speas 1999); Russian does not have sequence of tense, Amharic does not have sequence of person (Schlenker 1999).

Indirect speech is explicitly introduced by a reporter, and in sentences of indirect speech there is a systematic ambiguity. The report may be a recoding by the reporter of all or part of what was said, or it may be precisely what was said. In the former case some responsibility for formulation and truth is due to the reporter rather than the person whose communication is reported. (7) gives a well-known example.

(7) Oedipus said that his mother was beautiful.

This sentence could be used to report an utterance of Oedipus in which the speaker identifies the person that Oedipus talked about as Oedipus's mother (the *de re* reading). It could also be used to report exactly an utterance of Oedipus, "My mother is beautiful" (the *de dicto* reading). The truth of the sentence depends both on the interpretation in question and what was said: (7) might be true on the *de re* but not on the *de dicto* reading.

When deictic features that are oriented to the reporter appear in the complement the reporter clearly has some responsibility, as in (8):

(8) a. John said that Mary was leaving tomorrow.
b. John said that Louise is pregnant.

In these sentences an important component of the complement clause meaning is ascribed to the reporter. In (8a), the complement has the deictic adverb *tomorrow*, oriented to the time of speech, while the verb of communication is past. This conveys that the reporter has recoded all or part of what was said: the day of leaving is located with respect to the time of speech rather than to the past time of John's utterance. In (8b) the complement has present tense while the verb of communication is past. The sentence conveys that John's utterance about Mary's pregnancy was made in the past; and that the reporter relates the pregnancy to the time of speech, since (8b) would be false if Mary were not pregnant at the time of the report. Examples like this are known as 'double-access' sentences; they are discussed by Ogihara (1996), Abusch (1997), Giorgi & Pianesi (2000).

Given the close relation between them, we might attempt to derive indirect speech from direct speech. However it has been shown that this approach cannot work. The argument turns on two points: first, not all indirect speech reports have plausible counterparts in expressions of direct

speech; and secondly, the ambiguity between *de re* and *de dicto* readings arises only for indirect speech (Partee 1973; Banfield 1982).

The third member of the category of communication is 'represented speech and thought'. Represented speech has some features of direct speech and some of indirect. It presents the syntax of actual speech, but tense and pronouns are shifted as in indirect speech; (9) illustrates:

(9) Represented speech and thought
1 Mrs Dalloway$_i$ said she$_i$ would buy the flowers herself$_i$. 2 For Lucy had her work cut out for her. 3 The doors would be taken off their hinges; Rumpelmayer's men were coming. 4 And then, thought Clarissa Dalloway, what a morning — fresh as if issued to children on a beach. (Virginia Woolf)

The first sentence of fragment (9) has shifted tense and person: what Mrs Dalloway actually said, presumably, would have been *I will buy the flowers myself.* The subsequent sentences represent Mrs Dalloway's thoughts.

The syntax preserves the locutions of speech and thought. Represented speech and thought is not syntactically embedded: for instance, (9) ends with an exclamation, a form which can appear only in main clauses (topicalized sentences, elliptical fragments and a few other constructions are similarly limited). Represented speech may have a 'discourse' parenthetical, as in *He would be late, John said* (Reinhart 1975). The tense-aspect forms that appear in represented speech are limited in some languages. In French, for instance, the *imparfait* — a past tense with the imperfective viewpoint — is usually found in represented speech (Banfield 1982:158). Imperfectives and statives often occur in expressions of subjectivity; see 2.2.3 below for discussion.

Represented speech usually appears in fiction, where it can play an important role in conveying the character of a protagonist. Banfield suggests that it represents the consciousness of the person whose thought is presented (1982:10). Represented speech is also known as the *style indirect libre* (free indirect style), narrated monologue, and *erlebte rede*. Literary studies include Cohn (1978), Chatman (1978), Genette (1980), and others; Jespersen (1924) and Banfield (1982) have a more linguistic orientation.

2.1.2 *Contents of mind*

This category comprises expressions of mental states (such as thoughts, beliefs, and attitudes) and other propositional expressions. Standard examples of sentences expressing mental state have verbs like *think* and *believe*; their

complements express the object of thought or belief, the content of mind. The complements are similar in form to those of indirect speech; rather than external events of communication, they report mental states as in (10).

(10) a. We thought that Bella was in New York.
b. Mary$_i$ believes that she$_i$ won the race.

(10a) is subjective, expressing thought. There is tense concord between the main and complement clauses. In (10b), which expresses a belief, the main and complement clauses have coreferential subjects.

This coreference relation between referents in clauses involving thought or communication is sometimes called logophoric. Logophoric pronouns convey coreference between the subject of main and complement clauses in the context of verbs of thought, belief, or communication. Some languages have particular logophoric pronoun forms; one of these would appear in the complement of (10b). The term 'logophoric' is due originally to Claude Hagège, who studied African languages such as Ewe, Mundang, Tuburi, and Ubangi languages.[5] Logophoric pronouns have also been identified in Igbo and Gokana (Hyman & Comrie 1981), and Mapun, a Chadic language (Frajzyngier 1985), among others. In some languages, it is only verbs of psychological state and verbs of perception that introduce logophoric pronouns; in others complementizers and the subjunctive may also introduce them (Stirling 1993: 259 et seq, Giorgi & Pianesi 2000). The existence of logophoric pronouns (and related forms, discussed in section 2.2 below) shows that linguistic coding of access to mind is an important feature of language.

Reports of belief and thought involve the mind and consciousness of the person holding the belief. In Japanese there are particular features of vocabulary and syntax which are used to convey access to mind (Kuroda 1972). For instance, the pronoun *zibun* in Japanese conveys access to consciousness in communicative contexts: the referent is aware of the propositional content of

5. Hagège (1974) coined the term 'logophoricity' in connection with the African languages Mundang, Tuburi, and Ewe; he defined it as "to designate a category of anaphoric pronouns, personal and possessive, which refer to the author of a discourse or to a participant whose thoughts are reported" (Stirling's translation 1993: 253).

The term logophoric is now widely used, often in reference to discourse-oriented anaphors, or anaphors that appear in point of view contexts (Reinhart & Reuland 1993: 671). The term is also used for all kinds of indications of the mind or perspective of a participant; see the references cited in the text.

the clause containing the pronoun (Hirose 2000:1646). There are also perspectival uses of *zibun*, see Section 2.2.2 below.

Expressions of thought and belief are reported, like indirect speech; their formulation can be due to the reporter in the same way. (11), for instance, is ambiguous:

(11) Mary believes that that fool Gwendolyn wants to take over the committee.

On one reading, the epithet *that fool* is due to Mary: it is part of Mary's belief. There is also a reading, perhaps more salient in an isolated sentence, in which the reporter designates Gwendolyn as a fool. The two interpretations are true under different circumstance so that one cannot be substituted for the other. This property is known as 'referential opacity'; it is typical of complements referring to propositions.

Thoughts and beliefs are due to the person who holds the proposition. Vendler notes that propositions belong to particular individuals and are thus limited by subjective factors; in particular, referential opacity reveals the subjectivity of propositions (1972:73, 81). (12) illustrates; b and c are due to Peterson (1997) and Asher (1993) respectively.

(12) a. It seems that Mary will win the race.
b. Mary's having refused the offer was unlikely.
c. Everything that John believes is true.

We ascribe responsibility for the propositions to the reporter in these sentences. Predicates which take propositions as complements include *seem*, *appear*, *believe*, *fear*, *hope*, *want*, *think*, *affirm*, *deny*, *unlikely*, *impossible*, *inconsistent*, *sure*, *true*, *be certain*, *propose*, and *hypothesize*. Sentences with predicates which refer to propositions can be identified linguistically; sentences which directly express propositions cannot be.

Questions, imperatives, and sentences with modal auxiliaries or adverbials are another source of subjectivity in sentences. These forms express Projective Propositions, a class noted by Asher (1993). Projective Propositions are unrealized, as in (13):

(13) a. Clean up your room!
b. Will he leave the room?
c. He may/might leave the room.
d. John will probably win the race.

Projective Propositions also appear as the complement of verbs expressing

unrealized notions (e.g. *wonder*, *want*, *desire*, *guess*, *command*, *plead*, *entreat*, *allow*, *permit*). Projective Propositions can be distinguished from other propositions by their behavior under quantification.[6]

Modals, questions, and other projective propositions are ascribed to the reporter when they appear in main clauses, as in (13). When they appear in the complement clause of a mental or projective propositional verb, they are ascribed to the human subject of the verb, as in (14).

(14) a. John thinks that Mary will probably win the race.
b. They asked/ordered Mary to clean up her room.

In (14a) the modal is due to John, not the reporter; in (14b) *they* gave the order.

2.1.3 *Evaluation*

This class of expressions comprises evaluation, emotional reactions, commentary, and evidentials. All are subjective since all involve mind. They are conveyed by predicates and adverbials. The person responsible or affected is not always made explicit, as (15) illustrates.

(15) a. That Mary won the race was surprising to her/everyone.
b. That Mary won the race was surprising.
c. Surprisingly, Mary won the race.
d. That Mary won the lottery was lucky for her.
e. That Mary won the race was lucky.
f. Luckily, Mary won the race.

The person affected is identified in the prepositional phrases of (15a, d). Without such phrases (15b, e) we ascribe the effect to the reporter or to

---

6. In the quantificational test, one constructs a sentence with predicates referring to abstract objects (facts, propositions, projective propositions). If there is no quantifier that can appear in such sentences with a truth value, then the predicates take arguments of distinct, incompatible types. This test distinguishes Propositions from Projective Propositions: # indicates lack of truth value.

(i) a. #John desires everything that Mary believes.
b. # Everything that Mary asks for is true.
c. John asks for something that Mary wants.
d. Everything that John believes is true.

In (ia) and (ib) the clauses refer to different types of entities: in the first clause, the complements refer to projective propositions while the complements of the second clauses refer to propositions. They are semantically anomalous. In contrast, (ic) and (id) both have complements referring to propositions and the sentences are semantically well-formed (Asher 1993: 33–34).

another person (the latter interpretation is plausible in contexts where other participants are mentioned or assumed). Adverbials of this class are due to the reporter (15c, f) when they appear in a main clause. Many evaluative expressions have related 'psychological' verbs, e.g. *surprise*, *frighten*, *annoy*, etc. I do not think that these verbs are subjective in the sense being developed here: they do not require that responsibility be ascribed to a mind. For instance, we can account for a sentence like *That John was early surprised Mary* adequately with the notion of a thematic role of experiencer.

Commentary and evidentials are also inherently subjective. They are expressed by adverbials or predicates, and imply a responsible mind:

(16) a. Frankly, Mary won the race.
b. Clearly, Mary won the race
c. Allegedly, Mary won the race.
d. It was alleged that Mary won the race.
e. It was alleged by many observers that Mary won the race.

The reporter is responsible for adverbials of commentary and for some of evidentiality (16a–b). The situation is different for *allegedly* which means that an allegation was made by a particular person or source other than the reporter (16c). The adverbials in these examples correspond to the 'highest' classes of adverbs distinguished in Cinque (1999): Speech Act adverbials (*frankly*, *honestly*), Evaluative adverbials (*unfortunately*), Evidential adverbials (*clearly*, *allegedly*). Cinque gives a syntactically-based account in which a syntactic projection is posited for each class.

Evidentials may take other forms as well. Cinque points out that, in addition to predicates and adverbials, many languages have particles and idiomatic expressions of surprise, approval, etc. (1999). Cinque gives as an example the expression *after all* as in the sentence *So he is coming after all!* (*despite our expectation to the contrary*). Like other lexical expressions, cases like this are beyond the scope of this discussion.

2.2 Perspectival

Sentences with a particular perspective depart from the standard understanding of perspective as neutral, objective. By convention we understand the perspective of a non-subjective sentence to be transparent — not filtered through a particular mind — unless the sentence contains information to the contrary.

The central examples of this category are reports of perception. The

reported perception may be direct or indirect. Perception necessarily involves the particular perspective of the perceiver: reports of perception are subjective since a situation is perceived from the perceiver's standpoint. I include in this category sentences which convey or suggest a particular perspective. Perspective is conveyed by expressions which may be oriented either to the reporter or to a participant, primarily reflexives not syntactically conditioned, deictics, NPs and PPs that involve direction or location.

2.2.1 *Perception*

Linguistic presentation of perception may be direct, indirect, or inferred. The most straightforward cases are reports of direct perception: a verb of seeing, hearing etc. introduces a complement which expresses the situation perceived, as in (17):

(17) a. John saw that the sun was shining.
b. John saw Mary walk to school.
c. John saw Mary walking to school.

The examples illustrate the three forms of perception verb complements in English: propositional, a 'bare' or 'naked' infinitive, or gerundive. First-person reports of perception are also subjective: they express the perspective of the reporter as a participant (including the 'unreliable narrator' of fiction).

Less direct but very clear are cases where a sentence with a perception verb precedes another sentence; the second is taken to express the percept of the perception verb's subject. Examples like this tend to occur in narrative contexts. (18) illustrates:

(18) a. John looked out the window. The children were building a sandcastle.
b. Gabriel smiled at the three syllables she had given his surname and glanced at her. She was a slim, growing girl, pale in complexion and with hay-colored hair.(James Joyce)

These and other examples are discussed extensively in Caenepeel (1989). The second sentence in each case conveys the percept implied in the first: they a 'perspectivally situated' in Caenepeel's terms. Such sentences either have the imperfective aspectual viewpoint (18a), or express a state (18b); both express unbounded situations, for slightly different reasons.[7]

7. Aspectual viewpoints focus all or part of a situation. Imperfectives focus part of a

These expressions of indirect perception are like represented speech in some ways: both offer access to the mind of a participant, often with shifted deictics. Stative sentences, and sentences with the imperfective viewpoint, are usually found in these contexts. The reason is partly semantic: both focus an internal interval of a situation and thus lend themselves to the perspective of a participant in the situation. In French, only the *imparfait* may appear in such contexts (Banfield 1982: 158).[8]

Perspectivally situated sentences must present unbounded situations, according to Caenepeel. As evidence she gives examples like (19); the second sentence expresses an event with the perfective viewpoint.

(19) John looked out the window. Mary arrived.

It is difficult to interpret the second sentence here as conveying John's percept.[9] The more natural interpretation is that the two events occurred in sequence; or perhaps at the same time. Caenepeel explains these facts by appealing to our concept of perception. Perception is instantaneous. Therefore one can perceive only a short segment of an unbounded situation, an instant; a bounded situation requires more than a single instant.[10]

---

situation, excluding endpoints; perfective viewpoints focus a situation in its entirety, including endpoints or implicit bounds for events. In Russian there is a pragmatic convention in which the focus of the perfective is taken to be on the completion of an event; this convention is overridden under certain circumstances (Smith 1997). The imperfective viewpoint in English is also known as the 'progressive'; it is conveyed by the auxiliary *be+ing*.

The unbounded interpretation arises for different reasons in imperfective and stative sentences. In the former, the viewpoint does not include endpoints; in the latter, the temporal schema of a state has no endpoints (Smith 1991). In Caenepeel's work the two are grouped together into a supercategory of 'stative'; see also Herweg (1991). I argue against the supercategory approach in Smith (1996, 1999).

8. The French system is more strongly codified than the English. The French *imparfait* is a past imperfective tense which is commonly used to express perspective.

The English stative and imperfective together correspond grammatically to the French *imparfait*. The *imparfait* can be used for statives and non-statives, wheres the English progressive is possible neutrally only with non-statives. English stative sentences have the simple, perfective verb form.

9. Caenepeel claims that when a sentence presents a bounded situation the perspectival interpretation is unlikely at best. She says that bounded events — events presented perfectively — are impossible or awkward as perceptual reports unless there is a contingency relation between the event and the focalizing sentence.

10. This is essentially the same as Kamp's account of why present perfective sentences cannot be used to express a bounded event. We conceive of communication as instanta-

This explanation suggests another set of cases: in the context of sentences that express continous perception, perspectivally situated sentences with the perfective viewpoint should be appropriate. I think it is possible to construct such examples, in which perception takes place over an interval. (20) illustrates; the perception verb is perfective in (20a), imperfective in (20b).

(20) a. John looked out the window. Mary threw the ball to Sue and Bill played in the sandbox. The neighbor's dog arrived and trotted back and forth.
b. John was looking out the window. Mary threw the ball to Sue and Bill played in the sandbox. The neighbor's dog arrived and trotted back and forth.

Both examples convey ongoing perception, in which situations unfold as John looks from the window. They differ slightly. In (20a) the verb of perception has an inchoative interpretation, suggesting the beginning of John's looking; in (20b) the looking is in progress. Context plays a role in the interpretation of examples involving perception. In narratives it is common for thoughts and perceptions to be represented, often with some ambiguity as to whether the perspective belongs to the reporter or a participant. Thus narrative contexts — the main topic of Caenepeel's discussion — lend themselves to the perspectivally situated interpretation.

There are also cases of inferred perception. Sometimes one infers that perception is involved from the developing situation, as (21) illustrates. As in the cases of indirect perception more than one sentence is required. The first example is (1b), repeated here as (21a); the second is from Dowty (1986).

(21) a. I sipped my drink and nodded. The pulse in his lean grey throat throbbed visibly and yet so slowly that it was hardly a pulse at all. An old man two-thirds dead and still determined to believe he could take it.
b. John entered the president's office. The clock ticked loudly.

In both fragments the first sentence sets up a participant in a situation. The sentence(s) following are taken to express the percepts and/or thoughts of that person. Fragments that lead to the inference of perception tend to appear in

neous; communication of a bounded event, even if instantaneous, requires at least one instant after the completion of the event.

narrative contexts, in which one expects to find expressions of participants' perceptions.

2.2.2 *Particular standpoints*

I now turn to a different set of cases, where a particular perspective is suggested but perception is not involved. Rather, a situation is presented from the standpoint of a given participant, or by the reporter as participant. The notion of standpoint has a literal basis in the participant's placement in the world. From a particular location, an observing individual sees things in a certain way: if I say that something is nearby or in the distance, it is because of my position in space. If I say that the bank is around the corner and you say it is across the street, we can both be right if we are standing in different places (Mitchell 1986: 1). The notion of standpoint can be extended to situations in which one talks as if one were in a location; and metaphorically to attitudes and views that are not grounded in space.

Perspectival examples have reflexive pronouns, deictics, the imperfective viewpoint, and other expressions that imply a particular standpoint rather than an objective stance. (22) illustrates with reflexives:

(22) a. This paper was written by Ann and myself.
b. They$_i$ heard the stories about themselves$_i$.
c. Mary$_i$ put the blanket over herself$_i$.

The particular perspective of the reflexives' antecedent is suggested in these sentences. In (22a) the speaker must be the antecedent of the reflexive, and is plausible as participant (Ross 1970).[11] In (22b–c) the sentence subjects *they* and *Mary* are the antecedents of the reflexives, and the reflexive suggests that the stories and blanket are located with respect to these people, that is, from their standpoint or perspective; (22b–c) are based on examples from Cantrall (1974) and Kuno (1987).

When they convey a particular perspective, reflexive pronouns are optional rather than syntactically conditioned. Syntactically conditioned reflexives are obligatory in certain contexts; they are defined in Government

11. This is one of a set of examples which led Ross (1970) to suggest that all sentences have a higher clause in underlying structure with a first-person pronoun and a verb of communication in the present tense. The overt reflexive pronoun would be coindexed with the covert first-person pronoun. Harris reached the same conclusion on a much broader basis (1982), as Bruce Nevin has pointed out to me.

Binding theory with the notions of c-command and locality. Very roughly, the Binding Theory requires that the antecedent c-commands the reflexive if it is within the domain of the relevant governing category (Chomsky 1981).[12] Reflexives that violate these conditions represent a choice of the anaphor rather than a pronoun. To see the contribution of the reflexive, compare the sentences of (23). Both are well-formed.

(23) a. John$_i$ pulled the toy toward him$_i$.
b. John$_i$ pulled the toy toward himself$_i$.

The choice of the reflexive makes a difference in interpretation. (23b) suggests the perspective of John as he pulls the toy, (23a) has no such suggestion.[13] I shall refer to these and similar cases as "perspectival".

According to Kuno (1987), these examples convey an "empathy perspective" such that the reporter takes the perspective of a particular participant. They need not involve access to the mind of the participant, though they may suggest it.

The antecedent of a perspectival reflexive may appear in an independent sentence earlier in the discourse, as in (24), cited by Baker (1995):[14]

(24) She was not immediately able to say anything, and even when her spirits were recovered, she debated for a short time on the answer it would be most proper to give. The real state of things between Willoughby and her sister were so little known to herself, that in endeavoring to explain it, she might be as liable to say too much as too little. (Jane Austen)

---

12. The Binding Theory as stated in Chomsky 1981 has been the subject of much critical comment. Reinhart and Reuland 1993 offer an extensive revision in the same general framework; a different approach is taken in Pollard and Sag 1992.

13. Cantrall was perhaps the first to note the perspectival use of the reflexive. Cantrall presents many examples, among them the following sentences. Cantrall asks us to imagine that they describe a photograph which portrays a group of standing women who have their backs to the camera:

a. The women$_i$ were standing in the background, with the children behind them$_i$.
b. The women$_i$ were standing in the background, with the children behind themselves$_i$.

In (a) the children are located from the perspective of the speaker; in (b) they are located from the perspective of the women. As Zribi-Hertz notes, the sentences provide empirical evidence that the reflexive is correlated with an 'internal' point of view — that of a discourse protagonist as opposed to the speaker (1989: 704).

14. Alternatively, one might follow Harris in analyzing reflexives (1982) as reductions of metalanguage assertions of sameness.

Reflexives that can contrast with pronouns are known as 'Long Distance Bound' (Zribi-Hertz 1989), 'Long Distance Reflexives' (Stirling 1993), 'Locally Free Reflexives' (Baker 1995). I will use the latter term, Locally Free Reflexives, LFR for short, since examples like (23) do not involve a long distance. Not all LFRs are perspectival: they may also be emphatic or intensive (Zribi-Hertz 1989; Reinhart & Reuland 1993; Baker 1995); such cases are beyond the scope of this discussion.

Perspectival LFRs are coreferential in a way similar to the logophoric pronouns mentioned above, and indeed are called logophoric by many scholars. The antecedent to a perspectival LFR is a referent whose perspective is being represented. Such reflexives tend to occur in just those contexts in which logophoric pronouns may occur. However, they do not necessarily convey access to the consciousness of the participant.[15] Perspectival LFRs have been identified in many other languages, among them Japanese (Kuno 1972), Scandinavian languages (Thráinsson 1976), and Italian (Giorgi 1984). Kuno (1987) offers a survey of logophoric phenomena across languages.

Possessive pronouns may also suggest the perspective of a participant. They can do so, in English at least, because of the limited resources of the language. Possessives have no counterpart to the reflexive: there is only one

---

15. Hirose gives example of the two uses of *zibun*. The logophoric involves access to consciousness, the perspectival does not; his term for the latter is 'point of view.' Hirose says that in the logophoric example (a) Kazuo is obviously aware that he is shy, because he says so. On other other hand in example (b), Kazuo does not have to be aware that the book he lost is the one he borrowed from his friend. This is shown [by the fact that] (c) is not contradictory (2000: 1646).

(i) a. *Kazuo wa zibun wa tereya da to itteiru*
K. TOP self TOP shy.person COP QUOT say-STAT
Kazuo$_i$ says that he$_i$ is shy.

b. *Kazuo wa zibun ga tomodati karita hon o nakusit*
K. TOP self NOM friend from borrowed book ACC lost
Kazuo$_i$ lost a book that he$_i$ borrowed from a friend.

c. *Kazuo wa zibun ga tomodati karita hon o nakusita ga, sono*
K. TOP self NOM friend from borrowed book ACC lost but that
hon ga tomodati kara karita mono da to wa kizuite-it-nai.
book NOM friend from borrowed thing COP QUOT TOP realize-STAT-NEG
Kazuo$_i$ lost a book that he$_i$ borrowed from a friend but he has not realized that the book is the one he borrowed from a friend.

According to Hirose, *zibun* in examples like (ib) and (ic) conveys 'point of view', whereas in (a) *zibun* is logophoric.

form of possessive pronoun. (The possessive often appears with *own* in such cases, e.g. *her own house*.) Therefore possessive pronouns have the potential either for a reflexive or a perspectival reflexive reading. (25) gives examples in which possessive pronouns suggest the perspective of the antecedent: (25a) is from Kuno (1987), (25b) from Hirose (2000).

(25) a. John criticized his brother.
b. Kazuo lost a book that he borrowed from a friend of his.

The perspectival reading of the possessives is only optional in these examples, I think. They can be read as simply giving information about the relationship of the participants. For the latter interpretation one might say that the possessive pronouns do not have a reflexive component.[16]

Kuno finds that the perspectival effect also occurs in other cases of directional relationships. For instance, the phrase *John's sister* suggests John's perspective whereas *Mary's brother* suggests the perspective of Mary. If this is correct a sentence like (26) would offer the possibility of two perspectives.

(26) John and his brother talked to Mary about her sister.

If (26) suggests the perspectives of both John and Mary, it does not I think require the ascription of responsibility to either of them. The perspectival effects are relatively weak, then. Recognizing differences in strength among the relevant examples, Kuno (1987) posits a continuum of 'degrees of empathy'. At the high end of the continuum the reporter totally identifies with a participant; in the middle the identification is partial; at the low end the reporter manifests a total lack of empathy with participants. On such a continuum the examples of (25) and (26) are toward the low end.

Perspective may also be suggested by adjectives or epithets that would be expected from the participant, e.g. *John talked to Mary about his beloved cat.* Another source of participant perspective is the deictic adverbial. Deictics strongly suggest the perspective of a participant, especially when anchored to a time other than the time of speech, as in (27b–c):

(27) a. Mary lost her watch 3 weeks ago.
b. Mary had lost her watch 3 weeks ago.
c. Mary packed her clothes. She would be leaving soon.

16. Kuno, among others, regards posssessive pronouns as ambiguous between a [+reflexive] and [–reflexive] feature (1987:81).

In (27a) the deictic *3 weeks ago* is anchored to the time of speech. The deictics in the other two examples, *3 weeks ago* and *soon*, are anchored to past times not given explicitly; they take the perspective of Mary and therefore suggest her mind.

The imperfective viewpoint is hospitable to particular perspectives, as we have seen in the focalizing sentences of perception in (18a). Since the imperfective focuses an internal interval of a situation its formal meaning is compatible with its use to suggest an experiencing mind and/or a particular perspective. Traditionally, imperfective viewpoints are said to involve an 'internal perspective', whereas the perfective is external (Comrie 1976). Shifted deictics are clear linguistic evidence for the internal interpretation. They are always good in imperfective sentences, but limited in perfectives. This is particularly clear in French: the *imparfait* past tense allows shifted deictics more freely than the perfective past tenses (Banfield 1982; Smith 1991).[15] Oppositions such as perfective and imperfective often have the pragmatic function of marking what is traditionally referred to as 'point of view' in narrative (Fleischman 1991:26).

I conclude this brief survey by observing that there is an additive effect in sentences with linguistic forms which suggest the perspective of a participant but do not require this interpretation. If there is one such form, the suggestion is weak: with two or more, the suggestion of a participant perspective becomes stronger. (28) illustrates. In these examples, consider whether the question should be ascribed to the reporter or to Mary in a narrative context. Recall that questions suggest subjectivity, since they belong to the class of projective propositions (section 2.3 above).

(28) a. Mary played in the sandbox. Was it going to rain?
b. Mary was playing in the sandbox. Was it going to rain?
c. Mary was playing in the sandbox with her brother. Was it going to rain?

In (28a) the direct question is the only subjective element; in (28b) the preceding sentence has the imperfective viewpoint, which as we have seen invites a subjective interpretation. The example with the strongest subjective interpretation is (28c), which also has a possessive phrase oriented to Mary.

2.3 Summary and comment

Sentences require interpretations of subjectivity when they express a point of view, or take a perspective on a situation. Among sentences that express a

point of view, communication sentences involve public events, whereas contents of mind and evaluation express private events or mental states such as thoughts, beliefs, attitudes, evidentials, comment. These sentences have recognizable classes of predicates and adverbials; they also have characteristic patterns of deictics and aspectual categories.

The perspectival category includes perceptual reports and indications that a situation is viewed from a particular standpoint. Perspective is conveyed by verbs of perception, reflexives that are not syntactically conditioned, deictics, and other expressions involving directionality.

In interpreting subjectivity we ascribe responsibility for a clause to the REPORTER or the SELF. There may be uncertainty and/or degrees of responsibility. Indirect speech complements, and certain perceptual reports, are ascribed to either the REPORTER or the SELF. Perspectival sentences suggest the perspective of a SELF in varying degrees; in the weaker cases, we may ascribe the perspective to the REPORTER.

3. The composite approach

In most of the cases discussed above more than one linguistic form contributes to subjectivity either in the expression of point of view, or of a particular perspective. To make this concrete I list the forms most often found, in alphabetical order:

(29) Linguistic forms contributing to subjectivity
Communication and consciousness verbs
Main clause constructions mirroring idiomatic thought, speech
Complementizers
Deictic adverbials: place, time
Direction and location PPS
Epithets
Evaluative verbs and adverbs, conjunctions (*yet*, *anyway*, *still*, *after all*, *but*)
Evidential adverbials: *evidently*, *possibly*, *frankly*
Imperfective aspectual viewpoint
Lexical: verb direction (*go* vs *come*); psych verbs, dative verbs, etc.
Projective propositions
Pronouns, reflexives, possessives
Propositional attitude complements

Stative sentences
Subjunctive

The list is not complete but it gives a sense of the many and varied forms involved.

The composite approach looks at several linguistic forms and constructs an interpretation. In previous work I have presented such an analysis for aspectual situation types, e.g. states and different types of events. The interpretation of a sentence as stative, telic, atelic, etc. arises from a composite of the verb, its arguments, and relevant adverbials (Smith 1991). The rules for aspectual situation types focus on three aspects of surface structure: syntax, e.g. verb complement relations; categorial information (adverbial, NP, PP, etc.); and features which encode such information as verb class, definiteness, ±directional (for PPs), ±completive (for adverbials). The composite analysis looks at multiple information sources and constructs an interpretation. The composite rules are stated in the framework of Discourse Representation Theory.

The interpretation of subjectivity can be made explicit with an extension of the Discourse Representation Theory (DR theory) framework. The theory develops a dynamic representation of the truth-conditional and conceptual meanings of sentences in discourse (Kamp 1981; Kamp & Reyle 1993). The individuals, situations, and times that a discourse introduces are represented as distinct entities in the Discourse Representation Structure. The roles of SELF and REPORTER must be added to the DR theory repertoire if we are convey the meanings of subjectivity. A detailed proposal to extend DR theory in this way is made by Stirling 1993; Sells (1989) also gives a proposal.[17] In this article I will confine myself to a sketch of compositional rules that can analyze subjectivity.

DR theory posits construction rules which lead to a Discourse Representation Structure (DRS) for a discourse. The construction rules apply to sentences and deliver an interpretation to be encoded in the DRS.

Construction rules can account for the interpretations of subjectivity discussed above. Input to the rules is a syntactically analyzed surface structure. The rules recognize the combinations of linguistic forms that trigger an

17. Sells 1987 proposes that three roles be recognized: SOURCE, the one who makes the report; SELF, the one whose mind is reported; PIVOT, the one from whose physical point of view a report is made. For arguments against this view, see Stirling 1993.

interpretation of subjectivity. The output governs the interpretation: it will set up two roles, the REPORTER and the SELF, and ascribe responsibility for a clause to one or both of them. The role of SELF is identified with a participant in the situation; the REPORTER is always the writer/utterer of the clause. I will not attempt to deal with uncertainty or degrees of responsibility here. Relevant factors include the syntactic relations between clauses, categorial information, and such features as 1st, 2nd or 3rd person of pronouns; tense values; proximal vs. distal deictics.

As illustration, I will sketch a tentative analysis of sentences of indirect speech. The composite of forms which convey indirect speech are a verb of communication; the complementizer *that*; in the complement, a non-first person pronoun, past tense in matrix and lower clause, and whether nonproximal deictics. Much of this information can be encoded with features, e.g. verb class V[com], tense [past], NP [–1st person], Adverb [–proximal]. I will state interpretation principles for two cases: the case where the REPORTER is clearly indicated; and the default case, where it is not. In the default case there is no formal indication that the reporter is responsible for the clause of indirect speech, though it is always a possibility. If tense, proximal deictics, or other forms directly indicate the reporter, then both REPORTER and SELF are responsible for the complement clause; I will refer to them as 'reporter-based'. (This statement is too simple since it doesn't account for epithets or lexical cues, e.g. adjectives such as *beloved*.)

The composite of linguistic forms which triggers the interpretation of indirect speech is sketched in (30). Syntactic surface structures are relatively simple in DR theory, with few functional categories. The NP subject must be [+human]; the verb must be a verb of communication [V_{com}]; the tenses of both main and subordinate clauses, past; the subject of the lower clause 3rd person; adverbs in the lower clause may be proximal or distal. (30) includes only these forms and features; the rule does not 'see' other forms in the sentence.

(30) Indirect speech
$S[_{NP1subj}$ [+hum] V_{com} [past] . . . [that $_{S}[_{NP2subj}$ [3p] [past] $_{Adv}[$ ±proximal]]]]

(31) Interpretation of indirect speech
  a. Default
     REPORTER ← Main clause, possibly complement clause
     SELF ← Complement clause
  b. Reporter-based forms in complement
     REPORTER ← Main clause, complement clause
     SELF ← Complement clause

The default principle holds unless the complement clause has material explicitly related to the REPORTER; in that case the second principle applies. If the main and complement clauses have coreferential subjects, the interpretation is logophoric; this would be stated as an additional principle for a language with logphoric pronoun forms.

I now show how the rules apply to actual examples, two sentences presenting indirect speech. The first is (6a), repeated here as (32a); (33b) is closely related.

(33) a. Mary told me yesterday at the station that she would meet me there.
  b. Mary told me yesterday at the station that she would meet me here.

The sentences differ only in the adverb of the lower clause: it is distal in (33a) proximal in (33b). (34) shows informally how the construction and interpretation rules apply to (33a and b).

(34) a. S[Mary V_{com} [past] . . . [that [$_{NP}$[she] [would:PAST] $_{Adv}$[–proximal]]]]
     REPORTER ← Main clause
     SELF = Mary ← Complement clause
  b. S[Mary V_{com} [past] . . . [that [$_{NP}$[she] [would:PAST] $_{Adv}$[+proximal]]]]
     REPORTER ← Main clause; complement clause
     SELF = Mary ← Complement clause

(34a) gives the default interpretation, (34b) the case where the reporter explicitly shares responsibility for the complement clause.

Similar composite rules can be stated for the other cases discussed above. Since the rules apply at the level of syntactic realization, they must be stated separately for each case. For instance, there are different rules for propositional, modal, and evaluative predicates when they appear as verbs and adverbs.

As a final example I show how when there is a propositional, modal, or evaluative adverb in the main clause of a sentence, responsibility is ascribed to the REPORTER. When such adverbials appear in an object complement *that*

clause of a verb of thought, belief, or attitude, however, responsibility is ascribed to the SELF. The class of adverbs is notated as Adv_P.

(35) Propositional, modal, evaluative adverbs
  a. S[. . . Adv_P . . .]
     REPORTER ← Adv
  b. S[. . . $NP+hum_{subj.}$ . . . V_{com} . . . S[that [. . . Adv_P . . .]]]
     SELF ← Adv

Rules for other cases of subjectivity will follow the same lines.

4. Conclusion

In this discussion I have considered a variety of subjective phenomena. I reserve the term 'point of view' for sentences which express communication, content of mind, or evaluation. I use the term 'perspectival' for sentences which express perception or otherwise suggest a particular perspective on a situation.

The domain for the linguistic expression of subjectivity is the sentence, although it is usually at the discourse level that it is recognized and interpreted.

A composite analysis was sketched for sentences which express subjectivity, using compositional rules in the framework of Discourse Representation Theory. The rules recognize the key forms or combinations of forms that trigger interpretations of subjectivity. They ascribe responsibility through two roles, REPORTER and SELF, which are identified with participants in the Discourse Representation Structure. For a more complete treatment, see Smith (2003).

References

Sources for non-constructed examples are as follows, identified by example numbers:
(1a) Dick Francis. *Banker*. New York: Fawcett. 1984, p 3.
(1b) Raymond Chandler. *The Long Goodbye*. Penguin Books. 1953, p 64.
(1c) Alan Ehrenhalt, "Hijacking the Rulebook". *The New York Times*, Dec 20, 1998.
(2) John D. Scott and Tony Pawson. "Cell Communication". *Scientific American* June 2000.
(3) Peter Robinson. *A Necessary End*. New York: Avon Books, 1989, p 182.
(9) Virginia Woolf. *Mrs. Dalloway*. New York: Harcourt, Brace, 1925 (1981), p 3.
(18b) James Joyce. "The Dead". In *Dubliners*. Penguin Books. (pub 1916) 1958, p 177.
(24) Jane Austen. *Sense and Sensibility*. Oxford University Press. p 173.

Abusch, Dorit. 1997. "Sequence of tense and temporal de re." *Linguistics and Philosophy* 20: 1–50.

Asher, Nicholas. 1993. *Reference to Abstract Objects in Discourse.* Dordrecht: Kluwer.

Baker, C.L. 1995. "Contrast, discourse prominence, and intensification, with special reference to locally free reflexives in British English." *Language* 71: 63–101.

Banfield, Ann. 1982. *Unspeakable Sentences: Narrative and Representation in the Language of Fiction.* Boston: Routledge & Kegan Paul.

Caenepeel, Mimo. 1989. *Aspect, Temporal Ordering, and Perspective in Narrative Fiction.* Ph.D. Dissertation, University of Edinburgh.

Cantrall, William. 1974. *Viewpoint, Reflexives, and the Nature of Nounphrases.* The Hague: Mouton.

Chatman, Seymour. 1978. *Story and Discourse: Narrative Structure in Fiction and Film.* Ithaca: Cornell University Press.

Chomsky, Noam. 1981. *Lectures on Government and Binding.* Dordrecht: Foris.

Cinque, Guglielmo. 1999. *Adverbs and Functional Heads: A cross-linguistic perspective*. Oxford: Oxford University Press.

Cohn, Dorritt. 1978.*Transparent Minds: Narrative Modes for Presenting Consciousness in Fiction.* Princeton: Princeton University Press.

Comrie, Bernard.1986.*Tense.* Cambridge, England: Cambridge University Press.

Dowty, David. 1986. "The effects of aspectual class on the temporal structure of discourse: Semantics or pragmatics?" *Linguistics and Philosophy* 5: 23–33.

Fleischman, Suzanne. 1991."Verb tense and point of view in narrative." *Discourse-Pragmatics and the Verb*, ed by S. Fleischman & L. Waugh, Routledge: London.

Frazyngier, Z. 1985. "Logophoric systems in Chadic." *Journal of African Languages and Linguistics* 7: 23–27.

Genette, Gérard. 1980. *Narrative discourse: An Essay in Method.* Translated by Jane E. Lewin. Ithaca: Cornell University Press.

Giorgi, Alessandra 1984. "Toward a theory of long distance anaphors: a GB approach." *The Linguistic Review* 3: 307–361.

Giorgi, Alessandra & Fabio Pianesi. 2000. "Sequence of tense phenomena in Italian." *Probus* 12: 1–32.

Hagège, Claude. 1974. "Les pronoms logophoriques." *Bulletin de la Societé de Linguistique de Paris* 69: 287–310.

Harris, Zellig S. 1982. *A Grammar of English on Mathematical Principles.* New York: John Wiley & Sons.

Herweg, M. 1991."Perfective and imperfective aspect and the theory of events and states." *Linguistics* 29: 969–1010.

Hirose, Yukio. 2000. "Public and private self as two aspects of the speaker: A contrastive study of Japanese and English." *Journal of Pragmatics* 32: 1623–1656.

Hyman, Larry & Bernard Comrie. 1981. "Logophoric reference in Gokana." *Journal of African Languages and Linguistics* 4: 19–37.

Jespersen, Otto. 1924.*The Philosophy of Grammar.* London: Allen & Unwin.

Kamp, Hans. 1981. "A theory of truth and semantic interpretation." *Formal Methods in the Study of Language*, ed by J. Groenendijk, T.M.V. Jannssen & M. Stokhof, Mathemati-

cal Centre Tract 135, Amsterdam.
Kamp, Hans & Christian Rohrer. 1983. "Tense in Texts." *Meaning, Use and Interprtation of Language*. ed by Bauerle, R., C. Schwarze, & A. von Stechow. Berlin: de Gruyter.
Kamp, Hans & Uwe Reyle. 1993. *From Discourse to Logic*. Dordrecht: Kluwer.
Kuno, Susumo. 1972. "Pronominalization, reflexivization, and direct discourse." *Linguistic Inquiry* 3: 161–195.
Kuno, Susumo. 1987. *Functional Syntax*. Chicago: University of Chicago Press.
Kuroda, Yuki. 1973. "Where epistemology, grammar and style meet: A case study from Japanese." *A Festchrift for Morris Halle*, ed by Anderson, S. & P. Kiparsky. New York: Holt, Rinehart & Winston.
Mitchell, Jonathan E. 1986. *The formal semantics of point of view*. Ph.D. dissertation, University of Massachusetts at Amherst.
Ogihara, Toshi. 1996. *Tense, Attitude, and Scope*. Dordrecht: Kluwer.
Partee, Barbara. 1973. "The syntax and semantics of quotation." *A Festchrift for Morris Halle*, ed by Anderson, S. & P. Kiparsky. New York: Holt, Rinehart & Winston.
Peterson, Philip. 1997. *Fact, Proposition, Event*. Dordrecht: Kluwer.
Pollard, Carl & Ivan Sag. 1992. "Anaphors in English and the scope of the binding theory." *Linguistic Inquiry* 12: 2651–2305.
Prince, Gerald. 1987. *Dictionary of Narratology*. Lincoln, Nebraska: University of Nebraska Press.
Reinhart, Tanya. 1975. "Whose main clause? Point of view in sentences with parentheticals". In S. Kuno (ed), *Harvard Studies in Syntax and Semantics*, No. 1.
Reinhart, Tanya & Eric Reuland. 1993. "Reflexivity." *Linguistic Inquiry* 24: 657–720.
Ross, John Robert. 1970. "On declarative sentences." *Readings in Transformational Grammar*, ed by Jacobs, R. & P. Rosenbaum Waltham, Mass: Ginn & Company.
Schlenker, Phillip. 1999. *Propositional attitude and indexicality: A cross categorial approach*. Ph.D. dissertation, MIT.
Sells, Peter. 1987. "Aspects of logophoricity." *Linguistic Inquiry* 18: 445–479.
Smith, Carlota S. 1991. *The Parameter of Aspect*. Dordrecht: Kluwer.
Smith, Carlota S. 1995. *The relation between aspectual viewpoint and situation type*. Linguistic Society of America address. Published electronically, Eric database.
Smith, Carlota S. 1999. "Activities: States or Events?" *Linguistics and Philosophy* 22: 479–508.
Smith, Carlotta S. 2003. *Discourse Modes*. Cambridge: Cambridge University Press.
Speas, Margaret A. 1999. "Person and point of view in Navajo." *Proceedings of the West Coast Conference on Formal Linguistics*. CSLI Publications, Stanford, California.
Stirling, Lesley. 1993. *Switch-Reference and Discourse Representation*. Cambridge, England: Cambridge University Press.
Thráinsson, H. 1976. "Reflexives and subjunctive in Icelandic." *Proceedings of the New England Linguistics Society* 6: 25–39.
Vendler, Zeno. 1972. *Res Cogitans: An Essay in Rational Psychology*. Ithaca: Cornell University Press.
Zribi-Hertz, Anne. 1989. "Anaphor binding and narrative point of view: English reflexive pronouns in sentence and discourse." *Language* 65: 695–727.

Part 3

Syntax and semantics

Chapter 9

Some new results on Transfer Grammar

Morris Salkoff
Centre National de la Recherche Scientifique, Paris

1. A comparative French–English grammar

More than forty years ago in his article "Transfer Grammar" (1954), Zellig Harris addressed the problem of comparing the grammars of two languages. He sketched a comparison of the phonemics, the morphology, and the syntax of Hebrew, Korean, and English. I shall be concerned here only with his proposal for comparing the syntactic structures and sentences of a pair of languages.

In order to compare the sentence structures of two languages, he introduced a transfer relation between "each sentence of A and its translation in B, or between each grammatical construction of A and its translation in B [. . .]" (1954: 152). Harris gave a small sampling of such transfer relations between a few of the grammatical structures of English, Hebrew, and Korean. Even this small sample showed that it would be interesting from a linguistic point of view to constitute a more complete transfer grammar of a pair of languages.

I have completed a detailed comparative French–English grammar (Salkoff 1999) as a necessary preliminary step for a program of French–English machine translation (MT). In this grammar, French syntactic structures are translated into their English equivalents. This procedure extends Harris's proposal for defining transfer relations between two languages by providing a wide coverage of French syntax. This work was carried out independently of the results in Harris's paper, which I had read many years before. It was only after the comparative grammar had been completed and published that I re-read his paper and realized that I had fleshed out his proposal for the pair French–English. In fact, the resulting comparative grammar is interesting for at least three reasons: as an independent linguistic study, for the teaching of translation between the two languages, and for research on the possibility of setting up a program for machine translation.

The method used to construct a comparative grammar can resolve certain problems of ambiguity encountered in research on MT. Harris mentions the two principal difficulties: words having multiple translations, and structures having multiple translations. Regarding the first problem, Harris notes:

> When one word [. . .] has two translations [. . .] we can consider that the starting material is not merely the word in question, but the two environmentally distinguished occurrences of the word plus its environment, and each of these then has only one translation. (Harris 1954: 151)

The environment in question is the context containing the ambiguous word, and this context is taken into account in the comparative grammar to yield unique translational equivalents for each occurrence of word-in-context. An example from the French–English grammar is the translation of *seul* as *mere* or *only*.

In the noun phrase *le seul Na dN*, consisting of a definite article followed by *seul* and an abstract noun, where *Na* = *pensée* (*thought*), *synthèse* (*synthesis*), etc., the right adjunct (modifier) of the abstract noun, noted *dN*, cannot be empty. The translation of *seul* in this case is either *mere* or *only*, but it is possible to formulate rules which give the correct translation of *seul* for every form that the noun phrase *le seul Na dN* can take. When *seul* modifies abstract nouns like *pensée*, *espoir* (*hope*), *idée* (*idea*), etc., *seul* translates as *mere* when this noun is modified by various sentential right adjuncts, in particular, a complement clause or an infinitive. This is also the translation when the right adjunct of the noun phrase contains the indefinite prepositional phrase *de cela* (*of that*), which can be concatenated with the definite article to yield the demonstrative *ce*:

(1) a. Complement clause: *La seule pensée que Max exprimera sa colère me semble inacceptable* → The *mere* thought that Max will express his anger seems unacceptable to me
b. Infinitive: *La seule pensée de partir me fait peur* → The *mere* thought of leaving frightens me
c. Reduced adjunct: *Cette seule pensée* (*me fait peur* + *m'intimide*) → The *mere* thought (frightens + intimidates) me

For all other right adjuncts, the translation is *only*:

(2) a. Relative clause: *La seule pensée que Max exprimera est celle-ci* → The *only* thought that Max will express is this one
b. Adjective: *La seule pensée négative est celle-ci* → The *only* negative thought is this one

c. Present participle: *Le seul espoir soutenant ces gens s'est évanoui* → The *only* hope supporting these people has faded
d. Past participle: *La seule solution proposée par l'orateur était bizarre* → The *only* solution proposed by the orator was strange

A somewhat more complicated instance of the translation of *seul* is the case of the right adjunct *de N* on the noun modified by *seul*. In this case too, the two translations of *seul* can be separated by an examination of the syntactic relations between the adjunct *de N* and the abstract noun *Na* modified by *seul*.

With regard to the second problem, that of structures having multiple translations, Harris writes:

> When we find that a structure in one language is translated into two or more structures in the other, [. . .], we try to sub-classify it into two or more structures, each of which will have only one translation. If the structure is in terms of classes, we may succeed in this by dividing a class into subclasses. If possible, we find some property that distinguishes these subclasses. (Harris 1954: 153)

Harris then gives a few English-Hebrew examples in which some syntactic property is used to distinguish the translation of a structure. In this short paper, Harris limited himself to the comparison of structures that could be sub-classified and separated on the basis of their syntax. The example of *seul* above illustrates this sort of syntactically based subcategorization in the French–English comparative grammar. Further work with these multiple translations, however, shows that such a separation on the basis of syntax alone can be carried out only to a limited extent.

When a grammatical structure in language A corresponds to more than one grammatical structure in language B, it frequently turns out that a semantic subclassification of the ambiguous structure in language A yields separate structures, each with its appropriate translation from A to B.[1] The comparative French–English grammar furnishes many examples of the kinds of semantic sub-classes needed to separate the translations of ambiguous structures.

Consider, for example, the translation of sentences containing the predicate *être à l'abri de*:

(3) a. *Max est à l'abri* [*des* (*regards indiscrets* + *ennuis financiers* + *orages* + *gens importuns*) + *d'*(*un accident* + *une erreur*)] → Max is sheltered

1. The definitions of these semantic sub-classes of the major grammatical classes must be operational, so that any two lexicographers using them to classify a set of words should come up with substantially the same lexicon.

from (indiscreet looks + financial worries + storms + importunate people + an accident + a mistake)

b. *Max est à l'abri d'* (*un mur* + *un arbre*) → Max is in the shelter of a (wall + tree).

When the object is an abstract noun like *looks*, *worries*, etc., as in (3a), the translation of *être à l'abri de* is *be sheltered from*; when it is a concrete noun, like *wall*, *tree*, as in (3b), the translation is *be in the shelter of*. These results can be schematized in terms of semantic sub-classes as follows:

3 c. *N0 être à l'abri de* (*Na* + *Nh*) → *N0* be sheltered from (worry + people)

d. *N0 être à l'abri de Nc* → *N0* be in the shelter of a (wall + tree)

The rough partition of the nouns into three gross semantic subclasses, viz., *Na* (abstract), *Nh* (human), and *Nc* (concrete), enables us to write a formal rule for separating the two translations of *N0 être à l'abri de N1*.

Harris refers to this difficulty indirectly. When speaking of the problem of separating multiple translations of an ambiguous word, he says (1954: 151):

> Or the determining environment [i.e., context] may be the presence in the same sentence or discourse of other words drawn from one part of the vocabulary rather than from another [. . .] In such cases, the instructions may have to call for a sampling of certain neighboring words (often from among the members of particular word classes only).

These particular word sets will most frequently be semantic sub-classes of the major grammatical classes, and it is this case that cropped up most frequently in the construction of the comparative French–English grammar.

In the next example, the presence of words drawn from a specific and limited subclass of nouns allows us to separate the translations of a prepositional phrase headed by *vis-à-vis de*. The prepositional phrase *vis-à-vis de N* must be divided into three sub-structures in order to separate the translations, which vary also with the syntactic function of the phrase. When the phrase is a right adjunct of a noun, the translations are *about N* and *towards N*:

(4) a. *vis-à-vis de Na* → about *Na*

b. *Ses sentiments vis-à-vis de* (*cette fraude* + *la mort de sa mère*) → his feelings about (that fraud + the death of his mother)

c. *vis-à-vis de Nh* → towards *Nh*

d. *Son attitude vis-à-vis de* (*lui* + *la police* + *les jeunes officiers*) → his attitude towards (him + the police + the young officers)

As the right adjunct of an adjective, the phrase translates as *towards N* or *on N*, depending on the sub-class of *N*:

(4) e. *Max est* (*méfiant* + *intolérant*) *vis-à-vis* (*de la sociologie* + *de ce médicament* + *de ces racistes*) → Max is (mistrustful + intolerant) towards (sociology + this medicine + these racists)

f. *Max est dépendant vis-à-vis de cet analgésique* → Max is dependent on this pain-killer

In two contexts of comparison, when a specific semantic sub-class or a small group of words is present, *vis-à-vis de* has two special translations. One is the particular context of financial discourse, in which various currencies are being compared. Then the noun of measure in this domain is a currency, and *vis-à-vis de* translates as *against*:

(5) a. *Le dollar* (*s'effrite* + *tient bien*) *vis-à-vis* (*des monnaies européens* + *du yen* + *de la livre*) → The dollar is (declining + holding well) against (European currencies + the yen + the pound)

The second context is one of comparison, and is signaled by one of a small group of words like *rien* (*nothing*), *nul* (*zero*), *assez grand* (*rather large*), etc.; in this case, *vis-à-vis de* translates as *next to* (or *besides*):

(5) b. *Mon savoir* (*n'est rien* + *est nul* + *est petit* + *est très moyen*) *vis-à-vis* (*du sien* + *de celui de mon frère*) → My knowledge is (nothing + zero + small + very average) next to (his + that of my brother)

All of these contexts can be described formally in terms of the syntactic structures of the string grammar of French (cf. Salkoff 1973). This description yields the syntactic schemata on the basis of which the entire comparative French–English grammar can then be constructed. The English equivalents for each French syntactic structure presenting some difficulty or ambiguity of translation can be found in the manner exemplified above.

When the multiple translations of a given schema cannot be separated by these methods, then some approximation must be used. Harris mentions this difficulty (1954: 156), but suggests no solution:

> But for many purposes, [. . .], a many-one relation from the native to the new language is no trouble at all. The only trouble lies in the fact that the reverse would be one-many (i.e., we would have several translations among which we could not choose).

This difficulty arises frequently in a comparative grammar, and an approxima-

tion is necessary. I have used principally a *passe-partout* (all-purpose) approximation which gives a single translation in the one-to-many case; this unique translation is chosen so that it is approximately equivalent to each of the possible translations.

Consider, for example, the following sentences containing the idiom *y être pour Adv* (*beaucoup*):

(6) a. *Max y est pour* (*beaucoup + un peu + rien + quelque chose*) *si Luc a* (*réussi + échoué dans*) *ses examens* → Max . . . if Luc has (succeeded + failed) in his exams

b. *Max y est pour* (*beaucoup +*) *dans le* (*succès + échec*) *de ce projet* → Max . . . with (the success + the failure) of this project

The sequences of dots in (6) mark the places where a translation of *y est pour Adv* must be placed. The difficulty here is that two translations are possible for the idiom *y est pour beaucoup*: one in the context of success, and another in the context of failure. The translations of these sentences are given in one bilingual French–English dictionary as follows. When the *if*-clause or the *with N* phrase refers to an undesirable result (for the subject), the translations in (6a, b) are given as:

(6) a′. Max is largely to blame (if Luc has failed in his exams + in the failure of this project).

When, on the other hand, they refer to a desirable result, the translations become

(6) b′. (A lot + a little + none + some) of the credit is Max's (if Luc has succeeded + in the success of project).

Now, there is no difficulty in obtaining *Max is Adv to blame* or *the credit is Max's* from *Max y est pour Adv*. However, setting up formal rules that can distinguish correctly, on the basis of the further context, between a desirable and an undesirable result in the *if*-clause or the *with N* phrase is surely a Herculean task, and perhaps an impossible one, depending as it does on the attitudes and prior knowledge of speaker and hearer (or writer and reader). It suffices to consider the number of intermediate cases possible between success and failure, any of which may require very complex and/or extensive tests of context in order to place them definitely as one or the other. The entire question can be sidestepped by using the following *passe-partout* translations:

(7) a. *Max y est pour* (*beaucoup* + *un peu* + *rien* + *quelque chose*) *si Luc a* (*réussi* + *échoué dans*) *ses examens* → Max has (a lot + a little + nothing + something) to do with it if Luc has (succeeded + failed) in his exams
b. *Max y est pour* (*beaucoup* +) *dans le* (*succès* + *échec*) *de ce projet* → Max has (a lot +..) to do with (the success + the failure) of this project

The neutral translation of *y être pour Adv* (*si S* + *dans N*) as "have something to do with (it if *S* + *N*)" leaves it to the reader to assess whether blame or credit is to be attributed, and on whom it falls.

The *passe-partout* approximation turns out to be very useful in many situations of one-to-many translations. For it is frequently (but not always) possible to find an approximation that can suggest each of the possible translations to the reader, who, of course, has no trouble in understanding which one is meant when read in context. This approximation is also useful when there are multiple English equivalents for a given French schema because the latter is vague. Consider the possible translations of the prepositional phrase *le long de N*:

(8) a. *L'eau coule* (*tout*) *le long de la gouttière* → Water flows (all the way) (along + down) the drainpipe
b. *marcher* (*tout*) *le long de la rivière* → walk (all the way) (along + up + down) the river
c. *grimper* (*tout*) *le long d'un mât* → climb (all the way) (*along + up) a mast

In these sentences, the preposition *le long de* can be translated in various ways — *along*, *down*, *up* — according to which direction is appropriate both for the principal verb preceding *le long de N* and the sub-class of *N*. This difficulty is easily resolved by the reader who understands the context, but it would be extremely difficult or impossible to formalize the contexts in which *le long de N* appears so as to make the correct choice of translation. The *passe-partout* translation (*tout*) *le long de N* → *the* (*whole*) *length of N* avoids this problem by leaving it to the reader to understand what direction is meant:

(8) a. Water flows the (whole) length of the drainpipe
b. walk the (whole) length of the river
c. climb the (whole) length of the mast

The comparative French–English grammar makes extensive use of this *passe-partout* approximation, as well as of some others. The relatively wide coverage of the syntactic structures of French thus obtained has allowed some new and unexpected results to emerge.

2. New results

As a result of achieving wide coverage of the syntactic structures of French, some entirely novel sub-classes have come to light that are relevant in a comparative grammar, but not in the grammar of either language considered independently of the other.

For example, consider the classes *Va*(Fr), *Vb*(Fr) and *Vd*(Fr) of French verbs, some of which can appear with *en train de*, and others not. The corresponding English equivalents, classified in the corresponding sub-classes *Va*(Eng), *Vb*(Eng) and *Vd*(Eng), either can or cannot take the progressive tense:

(9) a. *Va*(Fr): *Max est en train de manger une pomme* → *Va*(Eng): Max is eating an apple
b. *Vb*(Fr): **Max est en train de vouloir sortir*; *Vb*(Eng): *Max is wanting to leave
c. *Vd*(Fr): *Max est en train de comprendre le problème* → *Vd*(Eng) Max is beginning to understand the problem; ??Max is understanding the problem

French verbs in *Va*(Fr) can take *en train de*, and the corresponding English verbs, in *Va*(Eng), can appear with the progressive tense *is -ing*. Similarly, there is a sub-class of French verbs, *Vb*(Fr), that cannot appear with *en train de*, e.g., *vouloir*; its English equivalent, *want*, is in the sub-class *Vb*(Eng), and cannot appear with *is -ing*. There is also a third sub-class of verbs, *Vd*, different from both *Va* and *Vb*, which can be defined as follows:

(10) Sbj *est en train de* Vd(Fr) → *Sbj* is beginning to *Vd*(Eng)

That is, the French verbs *Vd*(Fr), which can appear with *en train de*, just as verbs classified in *Va* can, are such that their English translation *Vd*(Eng) is a verb that does not usually appear in the progressive tense. Hence the translation of the nuance associated with French *en train de*, which is usually expressed (for verbs in the sub-class *Va*) by the English progressive tense, cannot be used here, but it can be approximated by *is beginning to*.

Neither of the sub-classes *Vd*(Fr) or *Vd*(Eng) is distinguished in the usual accounts of French or English grammar, since these sub-classes are defined by the requirements of a French to English translation. We can see this from the following considerations. Although the verbs in the sub-class *Vd*(Fr) can appear with *en train de*, they cannot be distinguished in a standard French grammar from verbs in the sub-class *Va*(Fr) solely on the basis of that criterion, since the verbs *Va*(Fr) can also appear with *en train de*. Rather, the syntactic behavior of *Vd*(Fr) is defined by the syntax of its English translation, as in (10). Similarly, in English grammar, *Vd*(Eng) is a subset of those English verbs that cannot appear in the progressive, i.e., a subset of the *Vb*(Eng) of (9b). Just which verbs of the class *Vb*(Eng) should be sub-classified in *Vd*(Eng) cannot be ascertained from the syntactic properties of the sub-class *Vb*(Eng). The verbs in *Vd*(Eng) are just those whose French equivalents can appear with *en train de*, as indicated in the translational equivalence (10). The sub-classes *Vd*(Fr) and *Vd*(Eng) are thus a result of a comparative French–English study, and not of any syntactic or semantic considerations usually found in French or English grammar.

A second result is furnished by those cases where a verb must be inserted into the English translation of a French schema. The verb in question has a particular syntactic relation to one of the nouns in the French schema. Consider the following sentence pairs:

(11) a. Max alluded to the crime
b. Max made an allusion to the crime

(12) a. *Max stabbed at solving the problem
b. Max made a stab at solving the problem

(13) a. John's denial contradicts Jim's tale
b. John's denial is in contradiction with Jim's tale

Clearly, there is some relationship between the (a) and (b) sentences of (11) and (13): in (11), the verb *allude* is nominalized by *make*, and its object *to NP* remains unchanged. Similarly in (13), *contradict* is nominalized by *be in*, and the preposition *with* is added. Sentence (12b) is constructed much like (11b): the verb *make* is followed by the nominalization *stab*; however sentence (12a) shows that there is no sentence equivalent to (11a) in this case, i.e., no sentence with a single verb replacing *make a stab*. There is as yet no established terminology for verbs such as *make*, *be Prep* which appear with these nominal-

izations;[2] their properties have been studied in detail only for French. Maurice Gross (1981) calls these verbs 'support' verbs, and I shall adopt his terminology. The support verb is designated as *Vsup*, and the noun associated with it is termed a predicate noun, which is noted as *Npred*. The sequence *Vsup Npred* is termed a support expression. The term predicate noun originates in the observation that the same kind of constraints are observed between *Npred* and the subject or complement of *Vsup Npred* as are observed between an ordinary (non-support) verb and its subject or complement.

Certain French noun phrases containing an *Npred* arise from the nominalization of a support expression, in which the support verb *Vsup* is deleted:

(14) *Max donne son acquiescement à ce projet* (Max gives his consent to this project) → *L'acquiescement de Max à ce projet* (Max's consent to this project)

In the example above, it was possible to translate *acquiescement* without reference to the support expression *donner son acquiescement* from which the nominalization is derived. However, some noun phrases of this type cannot be translated correctly without the re-insertion of the deleted *Vsup* into the English translation. This is the case for *entorse*, *précisions*, and *asile*:

(15) a. *l'entorse de Max aux règles* ↛ *Max's violence to the rules; → the violence *that Max did* to the rules;
b. *Les précisions du Premier Ministre au journaliste étaient nécessaires* ↛ *The full particulars of the Prime Minister to the journalist were necessary;
→ The full particulars *that the Prime Minister* (*gave* + *supplied*) to the journalist were necessary
c. *l'asile de la France aux réfugiés vietnamiens* ↛ *France's asylum to the Vietnamese refugees; → the asylum *that France* (*gave* + *granted*) to the Vietnamese refugees

In order to note this requirement in the transfer lexicon, a sub-class of French

2. Harris called verbs like *make* in (11) 'operator' verbs (1964: section 2). Some German linguists call these verbs 'Funktionsverben'; they have also been termed 'semantically empty' verbs. R. Cattell (1984) calls such a verb followed by a nominalization a 'complex predicate', and mentions Jespersen's term (MEG, 1965[1909–1949] VI:117) 'light' verb. Jespersen's terminology seems to be gradually coming into use. Grimshaw & Mester (1988) study the Japanese light verb *suru*, but neither this study nor that of R. Cattell is carried out systematically over the entire Japanese or English lexicon of verbs and nominalizations.

Npred like *entorse*, *précisions*, *asile* must be created, whose definition is that their English translation requires the insertion of the English equivalent of the deleted French *Vsup*. In (15a), the latter is *faire*, which translates here as *do*; in (15b, c) it is *donner* that must be inserted, and it translates as *give*. This subclass appears nowhere else in French grammar, for the only difference between *acquiescement*, which does not require the re-insertion of its *Vsup*, and *entorse*, which does, is their behavior under translation into English.

The foregoing examples give the reader an idea of the importance of the availability of a detailed French grammar for the construction of a comparative French–English grammar (and, *mutatis mutandis*, for any pair of languages). The possibility of separating multiple translations of words, or of structures, depends in large part on having a detailed syntactic and semantic description of the contexts in which these words and structures are embedded. Only then can the contexts be distinguished and the correct translation furnished. A French grammar is required with sufficiently large coverage to provide a framework in which such a description can be made. On the basis of such a French grammar, the French schemata can be described in enough detail that English equivalents for them, or satisfactory approximations of the kind discussed above, can be found.

The principal elements are now in hand, with which it is possible to construct a program of French–English translation. A detailed French grammar is available, and has been incorporated into a program of automatic syntactic analysis (Salkoff 1973, 1979); and a comparative French–English grammar has been constructed, which can be coupled with the decompositions provided by the syntactic analyzer to produce the desired translations. Both a syntactic analysis of the kind provided by the French string grammar, and a comparative grammar of the kind developed here are necessary preliminaries to the writing of an MT program between any two languages.

Harris was aware of the need for both of these components, when in "Transfer Grammar" he referred to MT and underlined the importance of syntactic analysis for any such program: "The analysis of a sentence into successively included constituents [. . .] is therefore necessary for any method of translation that is to be reducible to mechanical procedures" (p. 152). And his postulating a transfer relation between the sentences of A and their translation in B implicitly recognized the importance of a comparative grammar for the setting up of "mechanical procedures".

Here, as in so many other places, Harris's precursor work points in the direction that future research must take.

References

Cattell, Ray. 1984. "Composite Predicates in English". *Syntax and Semantics* 17. New York: Academic Press.

Grimshaw, Jane & Armin Mester. 1988. "Light verbs and Theta-Marking". *Linguistic Inquiry* 19: 2.205–232.

Gross, Maurice. 1981. "Les bases empiriques de la notion de prédicat sémantique". *Langages* 63. Paris: Larousse.

Harris, Zellig S. 1954. "Transfer Grammar". IJAL 20: 4.259–270. (Repr. in Harris 1970: 139–157. Page references in the text are to the reprint.)

Harris, Zellig S. 1964. "The Elementary Transformations". (= *Transformations and Discourse Analysis Papers*, No. 54.) Philadelphia: University of Pennsylvania. Excerpted in Harris (1970: 482–532).

Harris, Zellig S. 1970. *Papers in Structural and Transformational Linguistics.* Ed. by Henry Hiż. Dordrecht: D. Reidel.

Jespersen, Otto. 1965[1909–1949]. *A Modern English Grammar on Historical Principles.* 7 vols. London: George Allan & Unwin.

Salkoff, Morris. 1973. *Une grammaire en chaîne du français.* Paris: Dunod.

Salkoff, Morris. 1979. *Analyse syntaxique du français.* Amsterdam & Philadelphia: John Benjamins.

Salkoff, Morris. 1999. *A French–English Grammar: A contrastive grammar on translational principles.* Amsterdam & Philadelphia: John Benjamins.

Chapter 10

Pseudoarguments and pseudocomplements

Pieter A. M. Seuren
Max Planck Institute for Psycholinguistics

To what extent, and in what sense, transformations hold meaning constant is a matter for investigation; but enough is known to make transformations a possible tool for reducing the complexity of sentences under semantically controlled conditions. (Zellig S. Harris 1957: 340)

1. The problem of predicate–argument structure

Many attempts have been made to define the argument structure of predicates on semantic grounds. On the whole, such attempts have only been partially successful. Sometimes it seems clear that a predicate is complementary in the sense that it needs an extra term for it to make sense. An example is the predicate *build*: one cannot build without building something, which makes it look reasonable that the predicate *build* requires, besides a subject term, also a direct object term. The same goes for predicates like *be called*. One cannot be a name-bearer without there being an actual name: *I am called* is nonsense, but *I am called Pieter* is true. And an adjective like *worth* says very little about an entity to which it is applied unless it is specified what or how much the entity is worth. Yet a similar adjective like *tall* can be used absolutely of an entity, as when we say that Mount Everest is tall, even though it is not specified how tall the entity is. Both worth and tallness require measures or degrees, but the adjectives *worth* and *tall* differ in their conditions of use in that the former cannot, but the latter can, be used absolutely, without the required complementary term. The problems quickly accumulate. Consider the predicate *sit*. One cannot sit without sitting on something. Yet *sit* is rightly considered to be an intransitive verb, and the thing one sits on is denoted not by an argument term but by an adjunct of place.

As will be made clear more explicitly below, the term *adjunct* is used here for the surface structure realization of what are considered to be *adverbial operators*

in the semantic structure of sentences. Semantically, a sentence is considered to consist of a *lexical matrix*, which contains the main predicate (verb) and its argument terms, in the scope of a number of *operators*. An operator is 'abstract' predicate (i.e. specified in the lexicon as belonging to a word class that cannot function as a surface predicate) which takes an S-structure as subject term. This S-structure is its *scope* (see (9) below). The semantic function of an operator is, roughly speaking, to impose restrictions on the range of situations for which its scope, which always includes the matrix-S, is to be taken to hold.

An operator is *adverbial* just in case its surface realization is an adverb(ial particle) or PrepPhrase. Languages differ greatly in the way operators are realized in surface structure. For example, the word *just* is considered an adjunct in the English sentence *John has just left*, but not in its French equivalent *Jean vient de partir*, where *vient de* is part of the auxiliary verbal complex. And within English itself, the word *necessarily* is an adjunct in, for example, *That isn't necessarily true*, but *have to* belongs to the auxiliary verbal complex in the synonymous *That doesn't have to be true*. It is assumed, in this analysis, that operators are marked for their surface category in the underlying semantic structure that is input to the grammatical transformations.[1]

Sometimes one finds an interesting tension between adjunct and argument status. Some English verbs have prepositional objects, which may become the subject of a corresponding passive sentence, as in (1a), but not in (1b):

(1) a. The matter was dealt with by the manager.
b. *That day was left on by the manager.

In (1a) the main verb is clearly *deal with*, where the grammar must ensure that the preposition *with* is somehow placed over the object term (see (12) below). This is a clear case of a verb having a lexically defined prepositional object, which fits in naturally with the fact that *deal with* allows for passivization, as in (1a),

1. Superficially, the notion of operator used here differs from that adopted by Harris in, for example, (Harris 1978, 1981:412–435, 1982). In fact, however, there is a great deal of similarity. Whereas Harris speaks of operators in a more general sense as functions from expressions to expressions, they are defined here in a more restricted way as functions from S-expressions to S-expressions. Remarkably, Harris (1978:12, 1981:404) comes to the conclusion that "the only word classes needed for arguments are *N* [i.e. nominal; PAMS] or *O* [i.e. operator, for McCawley-Seuren: embedded S; PAMS]". This is precisely the conclusion reached by McCawley (1972:516–517) and Seuren (1985:113, 1996:25) with respect to the syntax of underlying semantic analyses.

with its stranded preposition *with*. However, matters are not always that perspicuous. Under certain conditions, English allows for passivization of prepositional objects whose argument status is less clear, as in (2a). The conditions under which such prepositional passivization is possible are not very well known, but it seems that a notion of cognitive 'distance' between what is specified in the PrepPhrase and what is specified in the main matrix clause plays a decisive role. Time adjuncts are apparently more distant from the matrix clause than adjuncts that have an intrinsic connection. Consider, for example, the following two sentences (the exclamation mark indicates pragmatic deviance):

(2) a. This bed has been slept in.
b. !This town has been slept in.

Clearly, the relation between sleeping and beds is as close as can be, much closer than that between sleeping and towns. Let us assume that this difference is expressed grammatically by the speaker's decision to treat the cognitively close adjunct *in the bed* as a 'low' operator, i.e. directly above the matrix-S. This very 'low' operator may be reinterpreted as being part of the matrix-S. In that case it is treated grammatically *as if it were a lexically defined prepositional object*, thus allowing for passivization. Such 'mongrel' argument terms may be called *pseudoarguments*. A cognitively more remote adjunct, however, like *in the town* in (2b) is treated as a proper adjunct of place that originates semantically as a 'higher' operator. Since a higher operator cannot be reinterpreted as if it were a lexically defined prepositional object and thus become a pseudoargument, it does not allow for passivization. The cognitive distance is then 'mirrored' in the grammatical distance. It will be shown below how this notion of grammatical distance can be made more precise.

Speakers determine the semantic rank, and thus the grammatical status, of phrases like *in the bed* and *in the town* on grounds of general knowledge and common sense. But a speaker may deviate from general common sense and decide to force a cognitive closeness between, for example, sleeping and towns. This has happened in (2b), which is not ungrammatical but only pragmatically deviant, insofar as the speaker has apparently decided that, in this case, the relation between sleeping and towns should be depicted as being as close as possible. Here the operator *in the town* has been placed just above the matrix-S, which has made it possible to treat it as a pseudoargument so that passivization is licensed. The result is a sentence that evokes the picture of a town whose streets have been messed up by, say, a garrison of soldiers that have spent the night there in tents or what not and have left behind a great deal of disorder.

In similar fashion, predicates or collocations that take on a metaphorical meaning may make a PrepPhrase change grammatical status. In (3a), for example, the verb *go* is interpreted literally, combined with a normal locative adverbial, which is too remote from the matrix clause to allow for passivization of the prepositional object. In (3b), however, the verbal collocation *go over* is naturally interpreted as a standard metaphor for dealing with an item on the agenda of a committee meeting. In that reading *go over* becomes a verb with a prepositional object as an object (pseudo-)argument that lends itself to passivization, just like *This matter has been dealt with.*

(3) a. !The bridge was gone over by the soldiers in five minutes.
b. The bridge was gone over by the committee in five minutes.

In general, it seems to be the case that we have at our disposal a number of reliable *necessary* conditions for argument status in relation to given predicates, but that we are still far removed from an adequate understanding of the *sufficient* conditions. That is, it looks as if all lexically defined argument positions express a function or role that is necessary for the predicate to depict a situation, but not all elements that are necessary for a predicate to do that are expressed as argument terms. Languages appear to have a certain freedom in this regard. If this is correct, it means that even if we cannot determine that certain forms of semantic content *must* be expressed as an argument term, we can decide that certain forms of semantic content *cannot* be expressed as an argument term. For if the semantic content in question is merely accidental to the relation expressed by the predicate, then, as a matter of principle, it cannot find expression as an argument term. This principle is important enough for it to be identified as the *Principle of the Exclusion of Accidentals* or PEA.

PEA seems to work reasonably well in the lexicon, especially with regard to nominal argument terms. If it turns out to be strictly observed by the languages of the world, it guarantees the exclusion of accidental semantic content from the argument frame of a predicate, but it does not guarantee the inclusion of nonaccidental semantic content. On the whole, this is what we find. Interestingly, however, PEA is sometimes violated, both in the lexically fixed argument frame of predicates and in the grammatical assignment of argument status. Such infractions of PEA do not seem to occur with nominal arguments but appear restricted to sentential argument terms. When this happens, we speak of *pseudocomplements*, a notion further elaborated in section 3 below.

However, before we pass on to pseudocomplements and the grammatical machinery required, let us play around a little more with nominal arguments. Verbs of giving, for example, require a beneficiary or else there can be no giving. Yet many languages do not express the beneficiary as an argument term (dative) but as a PrepPhrase or by means of a serial verb construction (SVC), as in some Creole languages (see (23d) below). The Romance languages have an obligatory PrepPhrase to express the dative, except with clitic pronouns, which occur in dative case. Some Creole languages use SVCs for datives, as has been said. English allows for either an argument term or a PrepPhrase, though the choice is not free for all verbs and those verbs that do have both constructions sometimes show subtle semantic differences. Green (1974: ch. 3) points out that verbs like *donate*, *give away* or *distribute*, for example, do not allow for an internal (argument term) dative but only for an external PrepPhrase with *to*:

(4) a. I donated/gave away/distributed old clothes to the Salvation Army.
b. *I donated/gave away/distributed the Salvation Army old clothes.

On the other hand, idiomatic expressions like *give a hug* only allow for internal datives:

(5) a. She gave her sister a hug.
b. *She gave a hug to her sister.

A semantic (though not truth-conditional) difference becomes manifest in cases like:

(6) a. She wrote a letter to the Pope.
b. She wrote the Pope a letter.

where (6b) evokes a relation of familiarity between the Pope and the letter writer, which is absent in (6a), where the relation is presented as formal.

The explanation is probably to be sought in the fact that entities referred to by means of a PrepPhrase tend to be accorded greater status and importance than those that are referred to by means of a canonical argument term. The difference stands out, for example, in otherwise symmetrical predicates like *shake hands with*. It is a widely known observation that (7a) will cause no surprise, whereas (7b) will make eyebrows go up owing to the importance the speaker implicitly accords to himself:

(7) a. I shook hands with the Pope.
b. The Pope shook hands with me.

Likewise, while the active sentence (8a) sounds normal, (8b) may provoke some puzzlement:

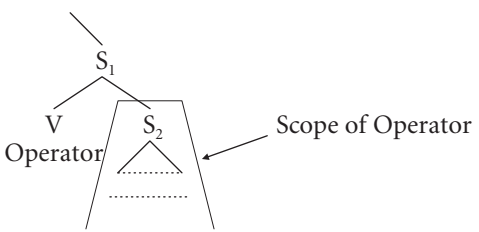
(8) a. John loves Africa.
b. Africa is loved by John.[2]

2. Some grammar

2.1 An overall view

A deeper explanation of this striking phenomenon is provided if it is assumed that PrepPhrases originate in the semantic analysis (SA) as operators that contain the matrix-S (S^M) in their scope, in the following way. We express the operator-scope relation in SA as a tree structure of the form (9), where S_1 may have a second NP or S argument following S_2, and where S_2 may again consist of a $_V$[Operator] and an embedded S, so that the structure is recursive. We assume a left-peripheral position for V (McCawley 1970), which is therefore followed by its argument terms. (Surface V-final languages are represented at SA-level with right-peripheral V.)

(9)



2. Harris attributes the difference illustrated in (8) to the past participle affix *-en*/*-ed* having "approximately the same semantic and descriptive relation to some such completive word as *state*" (1981:401) and to "the fact that the likelihood of making a passive for particular words depends on the likelihood of having an operator such as *state*" (1981:433). That is, Harris treats the subject term in sentences whose main predicate contains a past participle as being in some way said to be in a state defined by the participial construction. I find it hard to follow Harris in this regard. My reluctance to accept his account is based on two grounds. First, the subject-predicate debate, which raged in linguistics from about 1850 till 1930 (Seuren 1998:120–133), has shown that the grammatical notion of subject cannot be

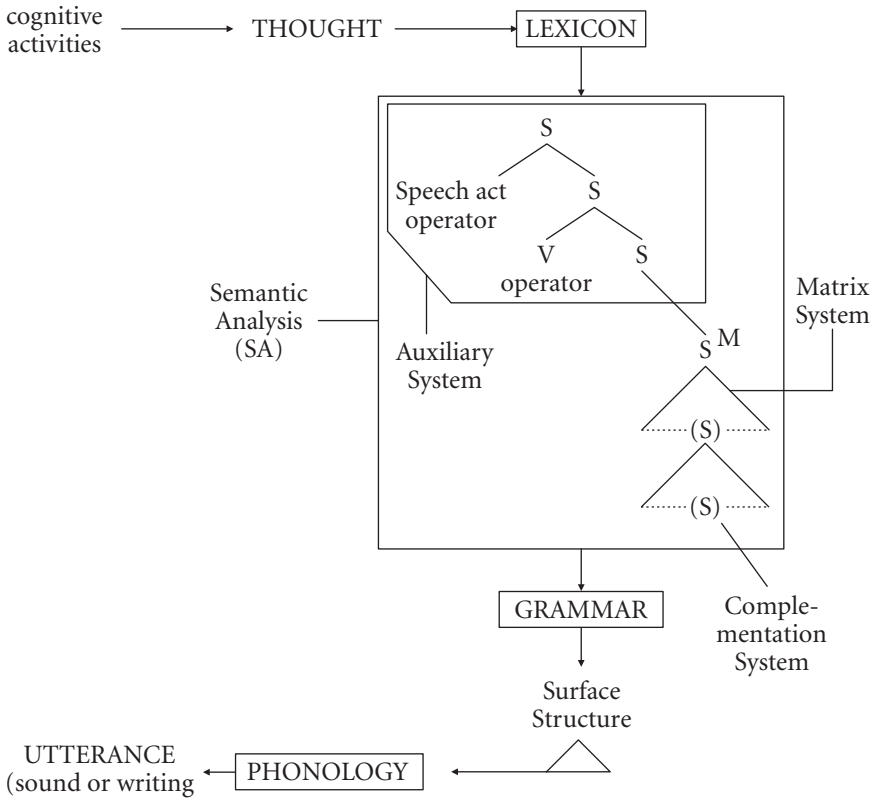


Figure 1. The architecture of a mediational top-down grammar

Languages have many different operators, all taking at least S^M in their scope, and thus occurring above S^M in the SA-tree. The upper part of the tree, which contains the operators, is called the *auxiliary system*, as opposed to the *Matrix System*, which looks at S^M, and the *Complementation System*, which deals with embedded subject and object clauses.

The overall architecture of a grammar of a natural language as envisaged here (for details see Seuren 1996) is shown in Figure 1. It follows the principle expressed in the quote from Harris (1957) heading this article. In this concept,

equated with any coherent semantic notion. Moreover, the account proposed here, based on the notion that entities mentioned in operators are assigned greater importance than entities mentioned as argument terms, has greater generality, as appears from exx. (6) and (7), which contain no passives or past participles (in (6) not even a subject term is involved).

it is assumed that the grammar mediates between thoughts and sound or writing by transforming semantically defined deep structures into surface structures. For that reason we speak of *Mediational Grammar*. The system is presented in a top-down fashion, since the bottom-up counterpart, the parser, is still in a rudimentary shape. A thought generated by a speaker (thinker) consists of a social commitment type plus a prelinguistic propositional structure that assigns a property to an entity. Given a thought, the speaker consults the lexicon of his language, where a search is carried out for the most appropriate lexical predicate available for the expression of the cognitive content contained in the thought. Some lexical items will be 'abstract' in that they do not appear as such in the surface structure for lack of a phonological specification. Other items are, as such, manifest in surface structure. The result will be an SA-structure, consisting of an auxiliary system containing operators of various kinds (including a speech act operator, not further discussed here), an S^M, and, optionally, one or more complement clauses.

In most languages, tense operators are obligatory (some languages, like Chinese, are perhaps best analysed without obligatory tenses). Other operators define modalities, place, circumstance, reason, duration, etc., mostly in the form of surface adverbials, i.e. adverbs or PrepPhrases. Quantifiers likewise function as operators, and so do negation, conjunction, and disjunction. The standard procedure for the operators of the auxiliary system is for them to be incorporated into S^M by means of the operation Lowering.

Operators are given the status of predicate (V), as they are always interpretable as expressions that assign a property to an entity, and because it simplifies and unifies the grammatical processing (McCawley 1972). The tense predicates and adverbials assign temporal and other properties to propositional objects.[3] Quantifiers assign quantitative higher order properties to (pairs of) sets.

Finite clauses standardly contain two tense operators, the finite tense operator V_{t1}, e.g. PAST, and the nonfinite tense operator V_{t2}, e.g. SIM (=simultaneous). The combination of PAST and SIM yields what is traditionally called

3. The ontological status of what are called here 'propositional objects' is a complex philosophical question. The term 'propositional object' is used here as equivalent to the extension of a proposition *p*, which is (nonstandardly) defined as the set of situations in which *p* is true, or the valuation space of *p* (see Seuren et al. 2001). Tense operators thus limit the valuation space of the argument proposition to certain time intervals, and adverbials impose other kinds of restrictions.

the simple past tense. V_{t1} can be filled by either PRES or PAST, V_{t2} by either SIM or PREC (=preceding). The four possible combinations of PRES/PAST and SIM/PREC yield the four tenses of the English tense system:

PRES + SIM → simple present (I walk)
PAST + SIM → simple past (I walked)
PRES + PREC → present perfect (I have walked)
PAST + PREC → pluperfect (I had walked)

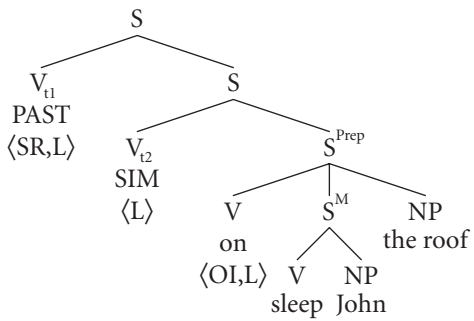
We shall, however, forgo a full discussion of the tense system, as it is not directly relevant in this context and would take up too much space. For some further discussion see Seuren (1996: 84–87).

2.2 Adjuncts, operators and (pseudo)terms

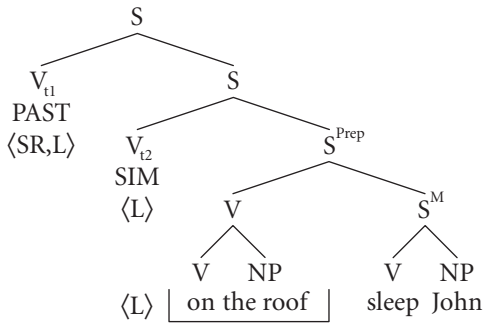
Let us now look in greater detail at prepositional operators. Consider sentence (10a), with its SA (10b) (the speech act operator has been left out):

(10) a. John slept on the roof.

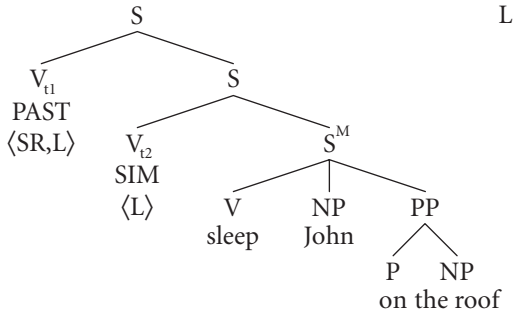
b.



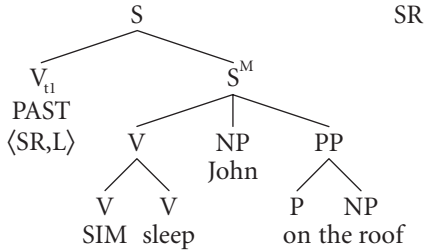
c.



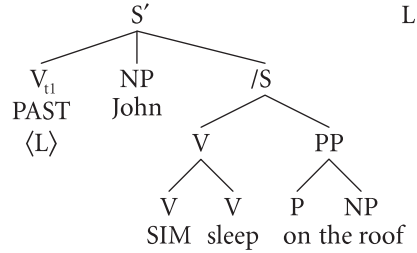
d. L →


S
V t1
PAST
⟨SR,L⟩
S
V t2
SIM
⟨L⟩
S M
V
sleep
NP
John
PP
P
NP
on the roof

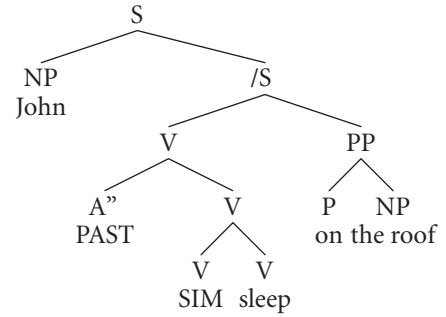
e. SR →


S
V t1
PAST
⟨SR,L⟩
S M
V
NP
John
PP
V
V
SIM
sleep
P
NP
on the roof

f. L →


S′
V t1
PAST
⟨L⟩
NP
John
/S
V
PP
V
V
SIM
sleep
P
NP
on the roof

g.


S
NP
John
/S
V
PP
A”
PAST
V
V
V
SIM
sleep
P
NP
on the roof

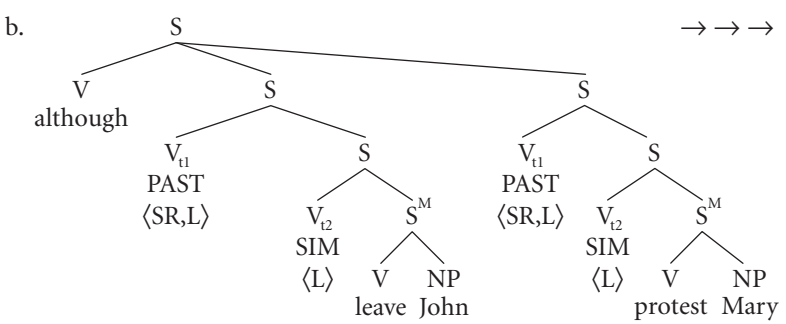
The SA (10b) is fed into the Grammar, where it first goes through the Cycle. The Cycle consists of rules that apply cyclically, i.e. it starts with the most deeply embedded S and works its way up through successive S-cycles until the top is reached. The rules to be applied, in so far as they are lexically defined for each predicate, are indicated in angled brackets for the predicate at each cycle. The first rule to be applied is Object Incorporation (OI). It takes the direct object of $S^{Prep}{}_{NP}$[the roof] and adjoins it to the first commanding predicate ${}_{V}$[on] to form a V-cluster, as in (10c). This V-cluster is then lowered into S^{M} in right-peripheral position by the rule Lowering (L), during which process the surface category labels PP (PrepPhrase) and P (Preposition) are assigned. The result is shown in (10d). Lowering of SIM on the S″-cycle as in (10e), and Subject Raising (SR) and Lowering on the S′-cycle as in (10f,g) then give the surface structure (10g).

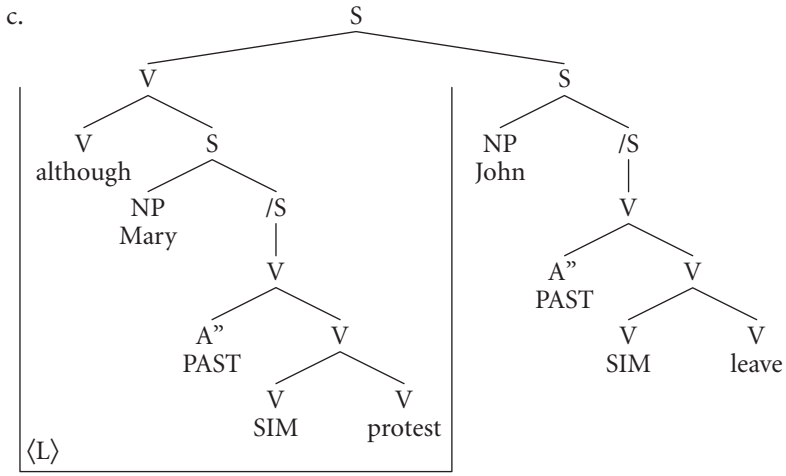
Note that S^{M} in (10e) is relabeled /S (i.e. incomplete S, or VP) in (10f). This is due to the general principle that any S that loses its subject-NP during the Cycle is demoted to /S (=VP). A second principle says that any S that loses its V during the Cycle is erased (no S without a V), all remaining material being united with the higher S in the order of occurrence. The postcyclic rules and the morphology will then produce the sentence (10a).

Adverbial subordinate clauses receive, in principle, the same treatment as PrepPhrases, although adverbial clauses tend to be placed higher up in the SA-tree. Consider sentence (11a), with the corresponding SA (11b). The conjunction predicate *although* takes two terms, a matrix subject term and a clausal object term, which again contains an S^{M}. Both are treated cyclically, and at the top cycle, where the two S nodes come together, the object clause is adjoined to ${}_{V}$[although] to form the V-cluster shown in (11c). After that, the V-cluster is lowered as a whole into the subject matrix S, where it can land in left or right peripheral position. (11a) shows the right-peripheral option.

(11) a. John left, although Mary protested.

b.





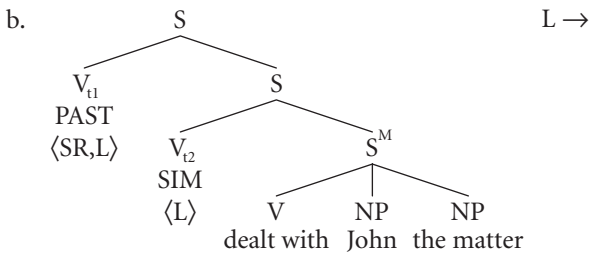
Right-peripheral Lowering of $_{V}[_{V}[\text{although}]_{S}[\text{Mary protested}]]$ leads to (11a), where *although* has been relabeled as a conjunction.

But let us now revert to the question that was raised at the end of Section 1: Why are entities mentioned in a semantic operator accorded greater status and importance than those that are referred to by means of a canonical argument term? Why should, for example, *to the Pope* in (6a) have a higher profile, in some not very clearly defined yet real sense, than *the Pope* in(6b)?

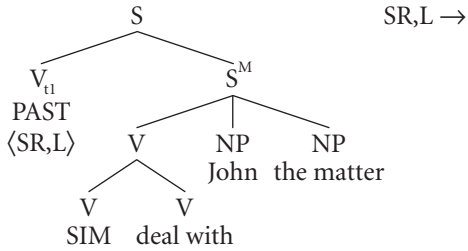
A similar question arises in connection with sentences like (2a,b), where passivization shows that the prepositional objects function as pseudo-arguments, i.e. as if they were canonical argument terms. Apparently, the process of being turned into a pseudoargument is constrained by the condition that the pseudoargument should have a relatively low profile with regard to the matrix-S. Once a nominal object term occurring in an operator has become a pseudo-object term, the further grammatical treatment is analogous to that of (12a), which contains a genuine prepositional object:

(12) a. John dealt with the matter.

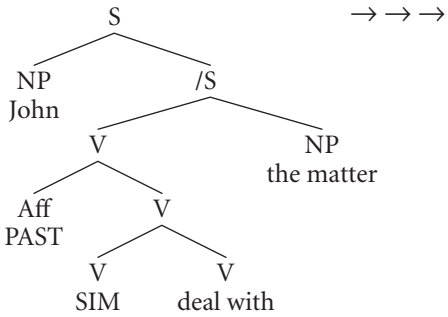
b.



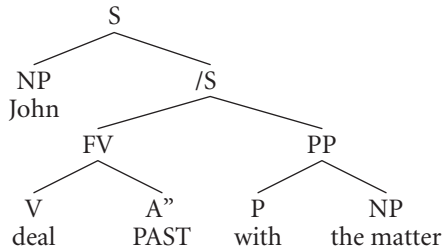
c.



d.



e.



In the opening sentence of section 2.1 we spoke of ‘a deeper explanation’. By this we meant that, as a matter of principle, a single argument term is semantically subordinate to a predicate–argument structure. This, one surmises, is so because a predicate–argument structure expresses a propositional thought that may be true or false and may be the object of a commitment or speech act operator, whereas an argument term is nothing but an element in a predicate–argument structure. When an S-structure S^n functions semantically as an argument term to a higher operator predicate in a higher S-structure S^{n-1}, then S^n is semantically subordinate to S^{n-1}. This may well explain why, as was observed by Steinthal (1855: 199), when we say *The patient slept well*, we usual-

ly mean to say that *the sleep of the patient was good.* And the negation *The patient did not sleep well* is normally interpreted as a negative comment on the quality of the patient's sleep, leaving the fact that the patient slept undenied. The manner adverb *well* is considered to represent an adverbial operator just above S^M, precisely like *to the Pope* in (6a), or *on the roof* in (10a), as shown in (10b,c). Likewise, a sentence like *Coffee grows in Africa* (Steinthal 1860: 102) will normally be used to say that the growth of coffee takes place in Africa, and its negation *Coffee does not grow in Africa* is a normal expression for the proposition that Africa is not where coffee grows.

For similar reasons, it is natural to assume that expressions like *by John* in (8b), or *with me* in (7b), or *to the Pope* in (6a) do not express terms but operators, hence predicates, and thus acquire a stronger 'profile'.[4] Note also that the negations of these sentences:

(13) a. Africa isn't loved by John.
b. The Pope didn't shake hands with me.
c. She didn't write a letter to the Pope.

are naturally interpreted as saying, respectively, that it is not by John that Africa is loved, that it is not me that the Pope shook hands with, and that it is not the Pope that she wrote a letter to (the negative sentence *She didn't write the Pope a letter* has that implication only with heavy accent on *Pope*).

2.3 Some notes on complementation

Let us now pass on to the Complementation System. In most cases the argument terms in S^M have the grammatical status of NP. However, the subject

4. Note that passivization of the prepositional object is excluded in (8b), which is already a passive, and in (7a,b), with the symmetrical predicate *shake hands* (symmetrical predicates preclude passivization), and also in (6a), whose passive would be *A letter was written to the Pope by her.* Note also that (8b) and (6a) may be taken to be instances of 'Argument Extraction', a process whereby, in the Semantic Analysis, argument material is lifted out of S^M and given the status of a (low) operator (Seuren 1996: 128–134), as a result of which the entity referred to takes on a higher profile. In English, this is a semantic option that makes for passives and external datives. Whether *with*-adjuncts to symmetrical verbs like *shake hands* (*with*), *meet* (*with*), or *agree* (*with*) can likewise be accounted for as the result of Argument Extraction from a coordinate structure subject term (*John and Harry shook hands/met/agreed*) is a matter for further investigation. In any case, it seems appropriate to assume that operators resulting from Argument Extraction cannot become pseudoarguments.

term and the direct object term , but never the indirect object term, may also be sentential. In that case the grammatical status is either S or NP-over-S, i.e. $_{NP}$[S]. Such embedded subject or object clauses are called *complement clauses*, and, as has been said, their grammatical treatment is the Complementation System of the language in question.

In the European languages at least, complement clauses occur in six possible forms, as S′ (i.e. with both tenses), S″ (with only V_{t2}), S^M, or as $_{NP}$[S′], $_{NP}$[S″], $_{NP}$[S^M]. Their standard surface realizations in English are shown in Figure 2.

S′ →	*that*-clause	$_{NP}$[S′] →	*that*-clause
S″ →	infinitival	$_{NP}$[S″] →	participial
S^M →	infinitival	$_{NP}$[S^M] →	participial

Figure 2. Six possible Complement-S-types

The English verb *believe*, for example, allows both for an embedded fully tensed S and for an embedded fully tensed $_{NP}$[S] in object position, both ending up as *that*-clauses. It also allows for an embedded S″ object clause, which ends up as an infinitival (*John believes Harry to be a linguist*).

The difference between embedded bare S and NP-over-S becomes grammatically manifest, for example, in S-anaphora: anaphoric S is *so*, as in *I believe so*, but anaphoric $_{NP}$[S] is *it*, as in *I believe it*. (Semantically, there is a difference in that the former is more appropriate for mundane belief-contents, whereas the latter is appropriate for major articles of faith.) For S′-embeddings it is likewise reflected in the grammatical status of the following sentences (Seuren 1996: 144–149):

(14) a. √Joe is likely to be ill.
b. √It is likely that Joe is ill.
c. √That Joe is ill is likely.

(15) a. √Joe seems to be ill.
b. √It seems that Joe is ill.
c. *That Joe is ill seems.

(16) a. √Joe tends to be ill.
b. *It tends that Joe is ill.
c. *That Joe is ill tends.

(17) a. *Joe follows to be ill.
b. √It follows that Joe is ill.
c. √That Joe is ill follows.

The differences in grammaticality are explained without further ado if the following complementation types are ascribed to the predicates in question for their sentential subject term:

likely	(Adj)	Subj: $_{NP}$[S′]/S″
seem	(Verb)	Subj: S′/S″
tend	(Verb)	Subj: S″
follow	(Verb)	Subj: $_{NP}$[S′]

Note, moreover, that the predicates *likely* and *follow*, but not the other two, allow for NP-subjects generally, as in *John's departure followed/was likely.* Assuming now that the cyclic transformational rule Subject Raising (SR) applies whenever the structural conditions are met (see below), the facts of (14)–(17) follow automatically.

The Complementation System is characterized by a small number of cyclic rules that form its core and are regularly encountered in the languages of the world. We shall mention three of them. The first is Subject Deletion (SD), or, as it was called in the early days of Transformational Grammar, Equi-NP-Deletion. This rule deletes the subject of an embedded object infinitival clause (S″ or S^M), which therefore becomes /S, under conditions of referential identity with a controlling NP (usually the subject) of the commanding S^M. More relevant in the present context, however, are the rules of Subject Raising (SR) and Predicate Raising (PR).

SR lifts the NP-subject of an embedded S″ or S^M, in either subject or object position, to the position of its own S, which becomes /S and is shifted one position to the right. There are thus two varieties: Subject-to-Subject Raising, as in, for example, (14)–(17), and Subject-to-Object Raising, as in *She wanted him to leave.* The facts are well-known for English, but SR is typical for the complementation systems of many other languages as well. Thus Portuguese, Russian, most Caribbean Creole languages, Latin, Ancient Greek, etc. are SR-languages (in traditional grammar the term *Accusativus/Nominativus-cum-Infinitivo* is used for what is now called SR). In fact, the group of languages whose Complementation System is characterized by SR is numerous enough to speak of SR-languages as a typological class.

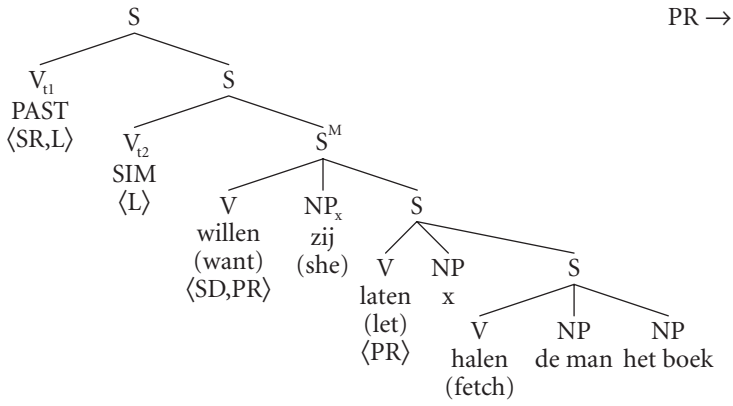
Note, however, that SR also occurs in the auxiliary system of NP-VP languages, where it is associated with V_{t1}, as shown in (10f) above. In general,

the cyclic rules may apply throughout the Cycle. Since, however, S^M is 'greedy' in that it incorporates elements from both the Auxiliary and the Complementation Systems, the raising rules are typical of the latter, whereas lowering is typical of the former. It remains to be seen to what extent the rule system can be simplified by uniting the raising and lowering processes into one superordinate scheme.

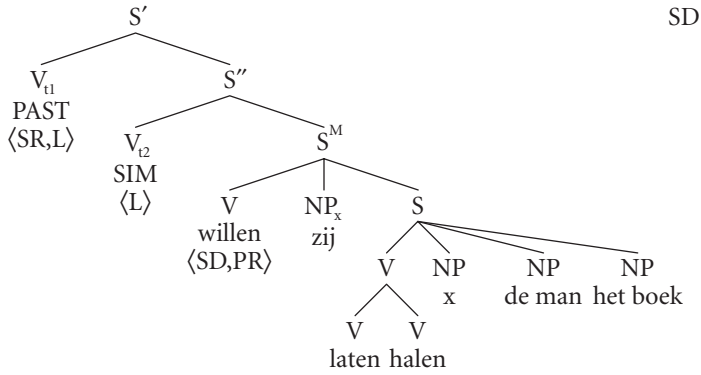
The third cyclic complementation rule to be mentioned is Predicate Raising (PR), untypical for English, but a dominant rule in many other languages, such as German, Dutch, Icelandic, French, Italian, Luiseño, Turkish, Japanese, to mention a few. Just as we think we may speak of a typological class of SR-languages, we may speak of a typological class of PR-languages. PR takes the V-constituent of the embedded clause and adjoins it to the V of the S where the raising takes place. Repeated application of PR leads to complex V-clusters, as shown in the Dutch example (18), (it is customary to present Dutch and German example sentences in the form of subordinate clauses, since these preserve the unity of the V-clusters, whereas the V-clusters are cut up into two parts in main clauses):

(18) a. . . . *omdat zij de man het boek wilde laten halen*
because she the man the book wanted let-INF fetch-INF
"because she wanted to let the man fetch the book"

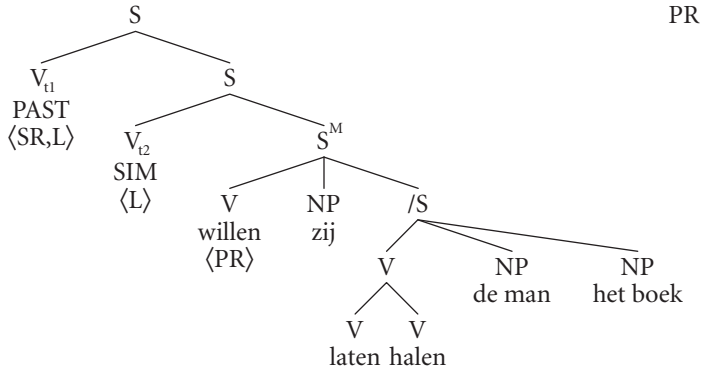
b. PR →



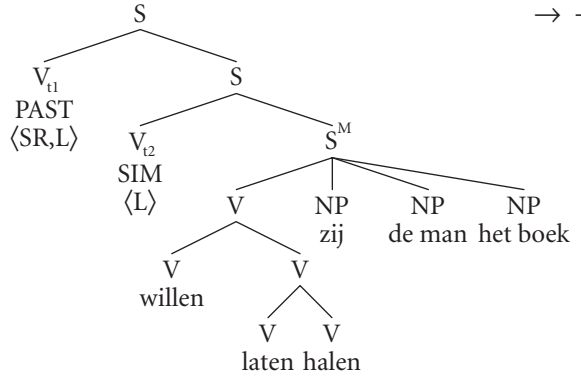
c. SD →


S′
V$_{t1}$
PAST
⟨SR,L⟩
S″
V$_{t2}$
SIM
⟨L⟩
S^{M}
V
willen
⟨SD,PR⟩
NP$_{x}$
zij
S
V
V
laten
V
halen
NP
x
NP
de man
NP
het boek

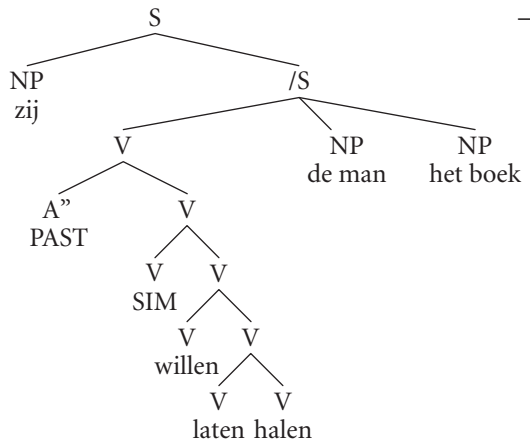
d. PR →


S
V$_{t1}$
PAST
⟨SR,L⟩
S
V$_{t2}$
SIM
⟨L⟩
S^{M}
V
willen
⟨PR⟩
NP
zij
/S
V
V
V
laten
halen
NP
de man
NP
het boek

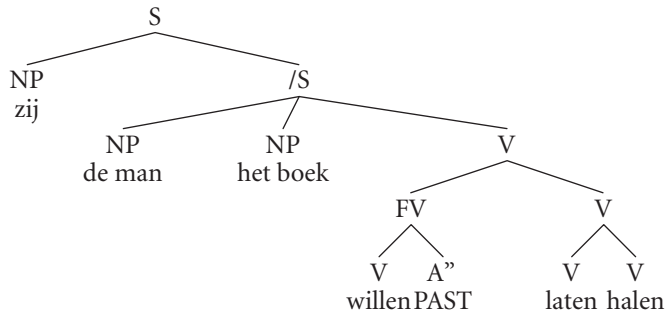
e. → → →


S
V$_{t1}$
PAST
⟨SR,L⟩
S
V$_{t2}$
SIM
⟨L⟩
S^{M}
V
V
willen
V
V
laten
halen
NP
zij
NP
de man
NP
het boek

f.



g.



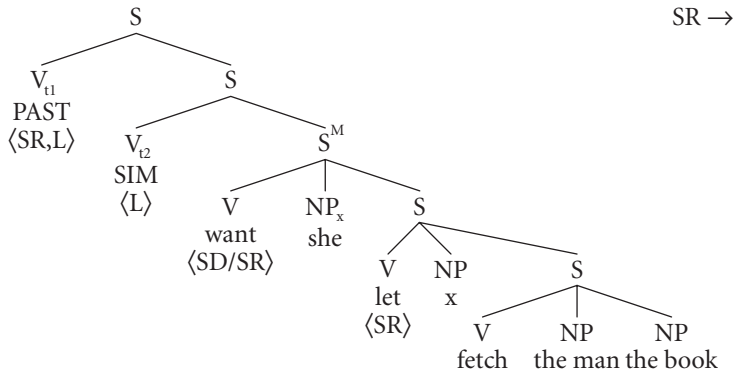
Cyclic application of PR to the SA (18b) gives (18c) (all raisings take right-attachment, resulting in a right-branching cluster). SD then deletes $_{NP}$[x] in (18c), as the higher subject is referentially identical to the lower subject (the identity is marked by the pronominal use of $_{NP}$[x]). The dominating S is turned into /S, as shown in (18d). Renewed PR yields (18e), where all complement-Ss have disappeared as they have been incorporated into S^{M}. The auxiliary system now gives (18f), which is where the Cycle ends. Postcyclic treatment deletes $_{V}$[SIM], unites $_{Aff}$[PAST] with $_{V}$[willen] into one finite verb form VF, and moves the entire V-cluster to the far right (as always in Dutch and German subordinate Ss). The result is (18g).

Note that, but for lexical differences and rule features, the SA (18b) is identical to its English counterpart (19b). The surface differences are caused exclusively by the rules associated with the various predicates. English *let* (in

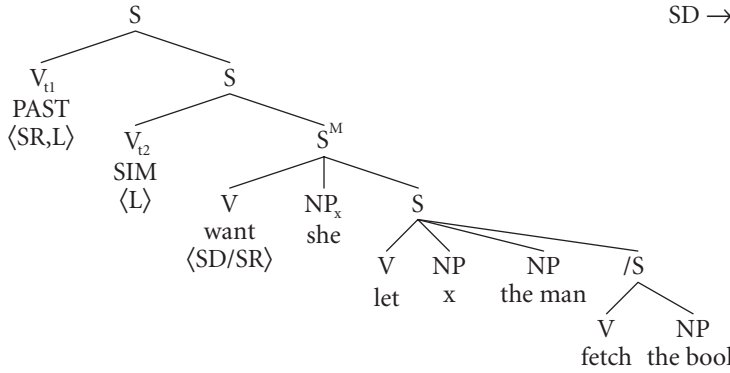
the raising, not the control, version) takes SR instead of PR, and English *want* takes SD or SR, the former when the higher and lower subjects are referentially identical, the latter when the two are referentially distinct (as in *She wanted John to fetch the book*). The derivation of (19a), in bare outline, is thus as shown in (19b–d), where we take for granted the insertion of the particle *to* with *let* and leave out the auxiliary part, which is self-evident:

(19) a. She wanted to let the man fetch the book.

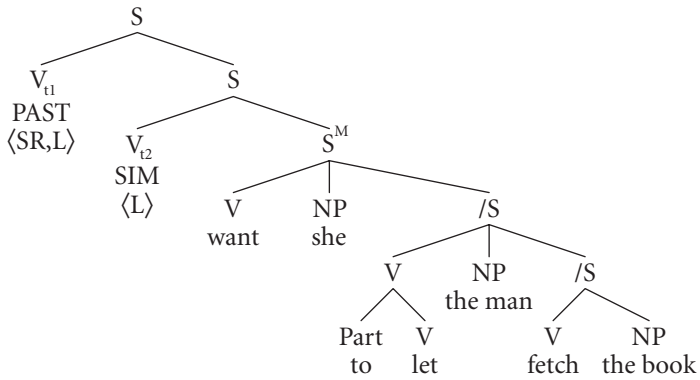
b. SR →



c. SD →



d.



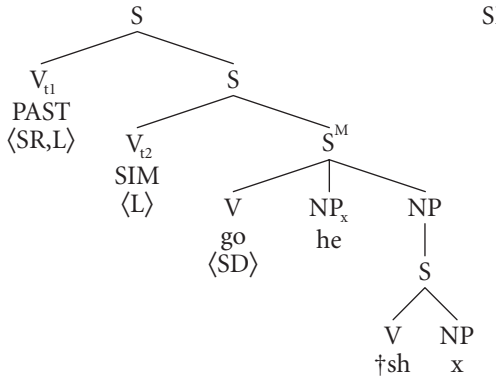
This much, incomplete and sketchy as it is, should suffice to enable one to capture the notion and the grammar of pseudocomplements, to which we now turn.

3. What is a pseudocomplement?

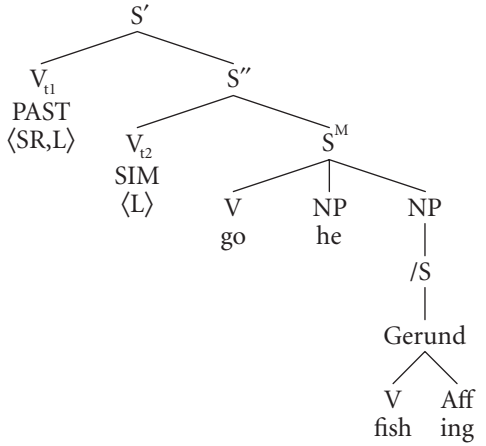
Let us return to the *Principle of the Exclusion of Accidentals* or PEA, mentioned above. It says that if the semantic relation of an entity *e* to the property expressed by the predicate *P* is merely accidental, then an NP referring to *e* cannot be an argument term of *P*. It was suggested that PEA is reliable for NP-arguments, and we will not try to undermine that suggestion. But we must make a serious reservation with regard to sentential arguments. For it does appear that predicates sometimes allow for object sentential complements that violate PEA. A simple case is the construction of the verb *go* with what is treated grammatically as an object-clause, as in (20).

(20) a. He went †shing.

b.



c.



The treatment of (20b) is transparent enough, if it is assumed that SD into an $_{NP}$[S] not only deletes the $_{NP}$[x] subject and turns the S into /S, but also adds an ING-affix and turns V into Gerund. In fact, there is no reason to treat (20a) in any way different from, for example, *He stopped fishing*, where *fishing* passes without any problem as an object clause.

In the SA (20b) *go* is given as a verb with an object $_{NP}$[S], even though what is expressed in this complement clause in no way satisfies a necessary condition for 'going' to take place. An event of going requires a subject term denoting the entity that goes, but it does not require a specification of the purpose of the event. Such accidental concomitants are normally expressed as

adverbial operators in the auxiliary system, and not as object clauses. They should be excluded by PEA, but, apparently, are not. The syntactic properties of *fishing* in *He went fishing* exactly match those of *fishing* in *He stopped fishing*. Treating *fishing* in (20a) as the surface representative of an adverbial operator in the auxiliary system would lead to considerable complications, making the grammatical system intolerably ad hoc, whereas treating it as a complement clause is syntactically without any problems (beyond those that exist anyway with regard to complement clauses).

The grammar thus puts the linguist under pressure to treat *fishing* in (20a) as a complement clause, even though the semantics of predicate–argument structures speaks against it. We note meanwhile that (20a) is not an isolated instance. Dutch offers many such examples, not only with the verb *gaan* ("go"), but also with verbs like *liggen* ("lie"), *lopen* ("walk"), *zitten* ("sit"), *staan* ("stand"). These are freely constructed with S-complements expressing an activity or a state, as in:

(21) a. *Jan ging vissen.*
Jan went fish-INF
"Jan went fishing."
b. *Jan lag/liep/zat/stond te dromen.*
Jan lay/walked/sat/stood to dream-INF
"Jan was dreaming (while lying down/walking/sitting/standing)."

In all these cases, the literal meaning of the main predicate has been 'bleached': the verbs in question no longer literally mean "go", "lie", "walk", "sit" or "stand", respectively, but rather indicate a state of being, with only weak connotations of going, lying, walking, sitting or standing. The important point, however, is that in (21a,b) the finite verb forms a V-cluster with the infinitive due to PR (we remember that Dutch is a PR-language). This appears, *inter alia*, from clauses like those in (22).

(22) . . . *dat Jan Marie de brief* [*liep te dicteren*]
that Jan Mary the letter [walked to dictate-INF]
"that Jan was dictating the letter to Mary (while walking)."

These are an exact match of cases where V-clusters arise as a result of PR applied to object clauses.

This shows with sufficient clarity that we have to do with clauses that are treated syntactically as object-complements, while, in virtue of PEA, the semantics of the predicates in question do not seem to allow for object clauses.

Such 'spurious' or 'mongrel' embedded object clauses are what we call here pseudocomplements. They express semantic content that is standardly expressed by means of an adverbial operator in the auxiliary system but has come to find a place as an object-complement. Accordingly, pseudocomplements express relations of purpose, concomitance, result, and the like.

The well-known serial verb constructions (SVCs), found, for example, in Chinese, Thai and many West-African and Creole languages, fit directly into this picture (Seuren 1990). Consider the following examples (the serial verbs are in bold):

(23) a. *Sùk ʔaw máy* ***maa*** *bâan.*
Sook take wood come house
"Sook brought the wood home." Thai (Schiller 1990)

b. *wǒ ná nèi-běn shū* ***gěi*** *le tā.*
I take DEM-CL book give PERF him
"I gave him the book." Chinese (Kortlandt 1998: 171)

c. *Kofi fringi a tiki* ***fadón naki*** *Amba.*
Kofi fling the stick fall hit Amba
"Kofi threw the stick at Amba." Sranan (Sebba 1987: 129)

d. *Kɔ̀kú pɔte kgab* ***ale*** *nã mâše.*
Koku bring crab go to market
"Koku brought a crab to the market." Haitian (Lefebvre 1986: 290)

e. *A man seri a buku* ***gi*** *a pikín.*
the man sell the book give the child
"The man sold the book to the child." Sranan

All such sentences are generated without a problem when it is assumed that the main verb in S^M has taken a pseudocomplement-S with an $_{NP}[x]$ subject term which is deleted by SD under referential control by a higher argument term. $_{NP}[x]$ is controlled by the higher object term in (23a, c, d), and by the higher subject in (23b, e). Note that (23c) contains two SVCs, a higher one with the verb *fadón* ("fall down"), which again embeds an SVC with the verb *naki* ("knock, hit"). Note also that the serial verb for "give" in (23b, e) fulfills the function of a prepositional dative: "to the child/him". Yet it does not originate as a higher operator but as an embedded pseudocomplement.

Interestingly, in many serializing languages, SVCs containing verbs of giving or going (as in (23a, b, d, e)) are in due course reanalysed as PrepPhrases

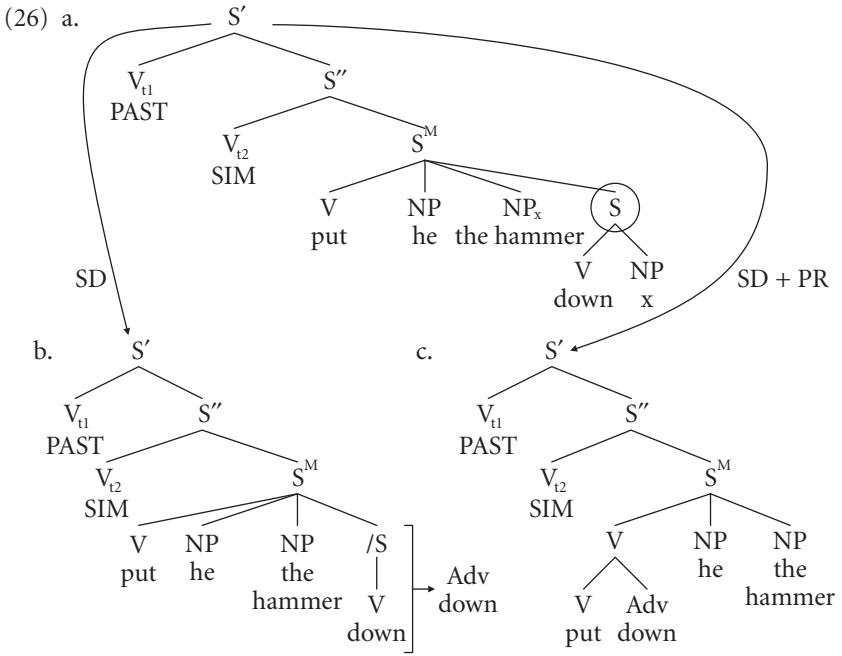
with the verb of giving or going reinterpreted as a dative or directional preposition, respectively. The fact that this does not happen with other verbs, such as those of falling or hitting, as in (23c), tells us something about the general property of language to encode certain semantic relations as prepositions but not others. Where to draw the line is a question that has, to my knowledge, not been investigated so far. In fact, the question has hardly arisen in linguistic theory, owing to the fact that it has not been customary to treat prepositions as semantic predicates.

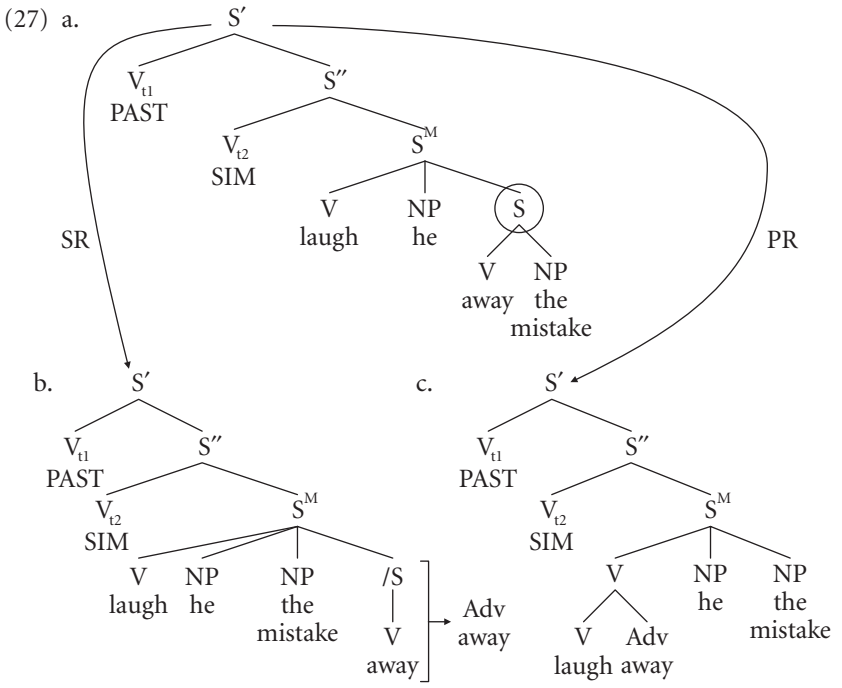
The decision to treat prepositions as predicates also sheds a new light on English resultative (quasi-)clauses that do not contain a verb but an adverb, PrepPhrase or an adjective, as is illustrated in (24) and (25):

(24) a. He put the hammer down.
b. He put down the hammer
c. He wiped the tears off his face.
d. He painted the door blue.
e. *He painted blue the door.

(25) a. He laughed the mistake away.
b. He laughed away the mistake.
c. He talked the invader out of the room.
d. He laughed himself silly
e. *He laughed silly himself.

While SVCs always instantiate a control structure, with the $_{NP}$[x] subject deleted by SD, English has both control and raising structures in nonverbal pseudocomplements. (24a–e) show control structures, while (25a–e) show raising structures. (26a) shows the putative SA for (24a, b); (27a) for (25a, b) (the pseudocomplement-Ss have been circled):


(26) a.
S′
V t1 PAST
S″
V t2 SIM
S M
V put
NP he
NP x
the hammer
S
V down
NP x
SD
SD + PR
b.
S′
V t1 PAST
S″
V t2 SIM
S M
V put
NP he
NP the hammer
/S
V down
Adv down
c.
S′
V t1 PAST
S″
V t2 SIM
S M
V
NP he
NP the hammer
V put
Adv down


(27) a.
S′
V t1 PAST
S″
V t2 SIM
S M
V laugh
NP he
S
V away
NP the mistake
SR
PR
b.
S′
V t1 PAST
S″
V t2 SIM
S M
V laugh
NP he
NP the mistake
/S
V away
Adv away
c.
S′
V t1 PAST
S″
V t2 SIM
S M
V
NP he
NP the mistake
V laugh
Adv away

(24a) is generated by the application of SD to the pseudocomplement; (24b) by the application of both SD and PR. It appears that PR, resulting in a complex V, is productively allowed in English only with pseudocomplements that have an adverbial predicate (like *down* or *away*). Otherwise, only SD is allowed, as in the control structures (24c, d), or only SR applies, as in the raising structures (25c, d). Thus, SR applies in the raising structure (27a), giving (27b), but PR may apply alternatively, giving (27c). Note that this account, if viable, replaces the traditional analysis in terms of the ad hoc rule of 'Particle Hopping' with one based on independently motivated rules and principles.

As predicted, (24c) is the result of simple SD, with $_{NP}$[x] deleted under control of the higher object term *the tears* in the pseudocomplement structure $_{S}$[$_{V}$[off] $_{NP}$[x] $_{NP}$[his face]], giving first $_{/S}$[$_{V}$[off] $_{NP}$[his face]], and then, by recategorization, $_{PrepPhr}$[$_{Prep}$[off] $_{NP}$[his face]]. Analogously for (24d), with the adjectival predicate *blue*: $_{S}$[$_{V}$[blue] $_{NP}$[x]] → $_{/S}$[$_{V}$[blue]] → $_{Adj}$[blue]. (24e) cannot be generated, as PR is not permitted with an adjectival predicate in the pseudocomplement.

Note that the Dutch equivalent of (24e), i.e. with PR and far-right movement of the V-cluster, is the only admissible form:

(28) . . . *dat hij de deur blauw verfde*
that he the door blue painted
"that he painted the door blue"

Here, the word group [*blauw verfde*] forms a V-cluster with the adjective *blauw* as a nonverbal element. This appears from the fact that nonverbal elements in Dutch V-clusters cannot occur at the bottom end of a V-cluster but must move upward, and that in moving upward they can take any position in the cluster, even at the top, but not outside the cluster. Thus Dutch has:

(29) a. . . . *dat hij de deur* V[*had moeten kunnen* ***blauw*** *verven*]
that he the door had must-INF can-INF blue paint-INF
"that he should have been able to paint the door blue"
b. *dat hij de deur V*[*had moeten* ***blauw*** *kunnen verven*]
c. *dat hij de deur V*[*had* ***blauw*** *moeten kunnen verven*]
d. *dat hij de deur V*[***blauw*** *had moeten kunnen verven*]

All of these mean the same. By analogy, one concludes that [*blauw verfde*] in (28) also forms a V-cluster, which can have resulted only from PR, the standard raising rule for Dutch.

3.1 Matrix Greed

Assuming that the analysis given above is viable or perhaps even correct, the question arises of the general rationale behind the whole system. This is a second order question, which can only be sensibly posed after a satisfactory analysis has been presented, but in that case it does present itself inevitably. This question is part of the general question of why humans do not speak in the language of Semantic Analysis, why the transformational machinery of the grammar appears to be a necessity. This question has so far not found a satisfactory answer, not least because we have no clear idea of the functional demands imposed on language for its smooth and proper functioning among speakers. I will, therefore, not try to answer that question here. But a less ambitious question can perhaps be formulated and answered.

If one tries to detect overall trends or tendencies in the grammatical machinery, one thing stands out clearly. The structural frame of a sentence is, on the whole, determined by the main lexical predicate and its argument terms, in other words, by what we have called S^M, the matrix-S, in the SA of sentences. Both the Auxiliary and the Complementation Systems shrink or disappear as they are, to some extent, swallowed up by S^M, which gets fattened. This appears to be a clear overall tendency in the grammatical systems of all languages. We call it Matrix Greed.

As we have seen, it is typical for the elements in the auxiliary system to be lowered into S^M. Likewise, in nonfinite complement-Ss, subject terms disappear or are incorporated into S^M by SR, or else the embedded V is adjoined to the higher V by PR, and the remaining material of the complement-S is amalgamated with S^M by unification. We have seen, moreover, that prepositional operators that are close to S^M in SA can, under certain conditions, be turned into pseudoargument terms, forming prepositional objects that are open to passivization. What we see in the case of pseudocomplements looks very similar: what should be a subordinate clause originating as an operator turns up, in some languages, as a spurious object clause, which is then processed as if it were a normal complement clause. Unlike the 'high' subordinate clause introduced by *although* in (11) above, pseudocomplements always seem to represent very 'low' operators, as close to S^M as is possible. This being so, one wonders whether the phenomena illustrated in section 1 above could not be seen as the *nominal* part of a more general manifestation of Matrix Greed. The principle of Matrix Greed would then be seen as allowing, under both language-specific and universal conditions, 'low' prepositional and clausal

operators to be incorporated into S^M as pseudoarguments. Pseudocomplements would then represent the *clausal* part of the same phenomenon.

References

Green, Georgia M. 1974. *Semantics and Syntactic Regularity.* Bloomington & London: Indiana University Press.

Harris, Zellig S. 1957. "Co-occurrence and transformation in linguistic structure". *Language* 33.3: 283–340. Repr. in (Harris 1981: 143–210).

Harris, Zellig S. 1978. "Grammar on mathematical principles". *Journal of Linguistics* 14.1: 1–20. (Also in Harris 1981: 392–411).

Harris, Zellig S. 1981. *Papers on Syntax.* Edited by Henry Hiż. Dordrecht: Reidel.

Harris, Zellig S. 1982. *Grammar of English on Mathematical Principles.* New York: Wiley.

Kortlandt, Frederik 1998. "Syntax and semantics in the history of Chinese". *Journal of Intercultural Studies*, 25: 167–176.

Lefebvre, C. 1986. "Relexification in Creole genesis revisited: the case of Haitian Creole". *Substrata versus Universals in Creole Genesis. Papers from the Amsterdam Creole Workshop*, *April 1985* ed. by P.C. Muysken & N. Smith. Amsterdam & Philadelphia: John Benjamins. 279–300.

McCawley, James D. 1970. "English as a VSO-language". *Language* 46.2: 286–299.

McCawley, James D. 1972. "A program for logic". *Semantics of Natural Language*, ed. by Donald Davidson & Gilbert Harman. Dordrecht: Reidel. 498–544.

Schiller, E. 1990. "The genesis of serial verb phrase constructions from an autolexical perspective". Paper presented at the Conference on Explanation in Historical Linguistics. University of Wisconsin, Milwaukee, April 1990. Ms.

Sebba, M. 1987. *The Syntax of Serial Verbs: An Investigation into Serialization in Sranan and other Languages.* Amsterdam & Philadelphia: John Benjamins.

Seuren, Pieter A.M. 1985. *Discourse Semantics.* Oxford: Blackwell.

Seuren, Pieter A.M. 1990. "Serial verb constructions." *When Verbs Collide: Papers from the 1990 Ohio State Mini-Conference on Serial Verbs.* Working Papers in Linguistics No. 39, ed. by B.D. Joseph & A.M. Zwicky. The Ohio State University Department of Linguistics, Dec. 1990. 14–33.

Seuren, Pieter A.M. 1996. *Semantic Syntax.* Oxford: Blackwell.

Seuren, Pieter A.M. 1998. *Western Linguistics. An Historical Introduction.* Oxford: Blackwell.

Seuren, Pieter A.M., Venanzio Capretta, & Herman Geuvers. 2001. "The logic and mathematics of occasion sentences". *Linguistics and Philosophy* 24.5: 531–595.

Steinthal, Heymann. 1855. *Grammatik, Logik und Psychologie, ihre Prinzipien und ihre Verhältnisse zueinander.* Berlin: Dümmler.

Steinthal, Heymann. 1860. *Charakteristik der hauptsächlichsten Typen des Sprachbaues.* Berlin: Dümmler.

Chapter 11

Verbs of a feather flock together II

The child's discovery of words and their meanings

Lila R. Gleitman
The University of Pennsylvania

> The whole schedule of procedures outlined in the following chapters, which is designed to begin with the raw data of speech and end with a statement of grammatical structure, is essentially a twice-made application of two major steps: the setting up of elements, and the statement of the distribution of these elements relative to each other. (Harris 1951:6)

During my years of graduate work in Linguistics at Penn I had the benefit of two profound and creative teachers, Zellig Harris and Henry Hoenigswald. Having passed the critical period for acquiring 30 or 40 ancient Indoeuropean dialects when I entered the field of Linguistics, I rapidly fell under the influence of Harris, whose thinking has guided the rest of my intellectual life. In light of that fact, I have been surprised, looking back over my own writings, to find that citations and references to Harris are conspicuously absent from most of them. In the context of the present volume, I have tried to think about why. The answer is that so much did Harris's approach to language get into my skin, become the sure and self-evident basis of my own thinking, as eventually to feel like my own quite clever inventions; that is, to lead to the well-known academic malady called Graduate Student Amnesia.

1. Acquiring the forms

The quotation from Harris that begins this chapter certainly sounds to modern-day psycholinguists like a prescription for language learning.[1] The infant's task

---

1. Harris himself did not particularly favor interpretation of his descriptive procedures as learning theories although he acknowledged that "the present operations of descriptive

in acquiring a language begins with the requirement to find and store recurrent patterns in the sound stream of input speech, and to classify these as linguistically significant elements which, in turn, are combined as higher-order elements. Thus the elements at each ascending level are created by the distributional analysis of elements at the level below. For instance, the infant learning English must discover that this language is built out of such sound segments as *t*, *p*, and *a*, and that these occur within such larger units as *tap*, *apt*, and *pat* but not *pta*. Then the child can begin to observe how these words, in turn, are distributed in sentences and in communicative contexts. These segmental choices and classifications are the primary data for the acquisition of the syntactic and semantic organization of the native language. Viewed from recent perspectives, such (roughly) bottom-up analyses of the speech stream as Harris's seem to describe the first two years of infant life in remarkably close detail. To me, these natural infant 'discovery procedures' for the categorial structure of the native tongue (of course, by creatures endowed by nature with Language-capital-L in advance) by successive distributional analyses strikingly resemble the Harrisian 'checking procedures.'[2] Some details of recent experimental findings documenting such procedures in infants follow.

---

linguistics can serve the further investigations which will obtain" such results (1951:375). These disavowals of course didn't stop his students from being inspired toward such extensions. It's probably correct also to suppose that Harris didn't view the machine analyses of language structure (the so-called TDAP parser designed for the Transformations and Discourse Analysis Project under Harris's aegis by A.K. Joshi, N. Sager, and a few others, tangentially including myself, and further developed by Sager and others at NYU) as anything like a real-time parser as it is thought about in the recent computational linguistic and psycholinguistic literature. Nevertheless this interpretation has been brought out clearly in a recent faithful reconstruction of this early work (Joshi 1999, and in Volume 2 of the present work, Chapter 5) and, in many ways, the original work with Harris has influenced Joshi's formal linguistic work on string adjunction and, later, tree adjunction.

2. By "endowed with language-capital-L" I mean that learners could not succeed if they were altogether openminded about the analyses (including the particular statistics) they will perform on the input corpus (Aslin, Saffran, & Newport 1998; Marcus 2000), about the concepts that languages encode (Baillargeon 1998; Spelke, Breinlinger, Macomber & Jacobson 1992), and particularly about how languages render predicate–argument structure (Chomsky 1981; Jackendoff 1983; Gleitman 1990). Sufficient demonstration in this last regard comes from examination of the 'home sign' and creolizing gestural communication systems developed by deaf children isolated both from signed and spoken languages (Feldman, Goldin-Meadow, & Gleitman 1978; Gleitman & Gleitman 1997; Goldin-Meadow & Mylander 1984; Newport 1990; Senghas, Coppola, Newport & Supalla 1997; Senghas &

1.1 The base units

Learning the forms and structures of language is grounded in human infants' inborn perceptual-phonetic capacities. Early in the first year of life they readily discriminate speech sounds that fall into different phonetic categories (e.g., discriminating the syllable /ba/ from the syllable /pa/, see Eimas 1975; Werker & Tees 1999), including sounds that do not occur in the language they are hearing from adults. In a few short months, infants redirect their attention to contrasts relevant to their native language phonology (Jusczyk 1997). A traditional linguistic assumption is that semantic learning could drive this reorganization: Infants could learn to ignore phonetic differences that do not signal meaning differences (e.g., Bloomfield 1933), using pairs of words that differ minimally in sound but contrast in meaning to determine the phonetic inventory (*bare*/*pare*, etc.). However, given the early narrowing of the phonemic distinctions recognized reliably by infants, there is every reason to believe that the migration of perception toward language-specific phonetic patterns is largely a matter of perceptual learning without early feedback from the semantic system — a distributional analysis for the recurrent forms much in the spirit of Harrisian checking procedures. In particular, consonant and vowel tokens in different languages differ in their distribution across various acoustic dimensions and help the child redistribute her attention within the phonetic perceptual space (Kuhl 1994; Jusczyk 1997). Distributional learning, in many of the senses laid out by Harris, is thus directly relevant to the early categorization of speech sounds.

As noted by Harris in the paragraph which heads this chapter, this twice-applied distributional analysis (the first to discover some set of elements, the second to ascertain their distribution in utterances) is repeated again and

---

Coppola, in press). These cases show that isolated infants can and will invent communication systems that look like the received languages, both in form and content. Harrisian distributional procedures are closely relevant to how the child discovers a language 'in the stimulus,' that is, how he or she acquires the instantiations of particular categories in the exposure language. These must be *learned* in the traditional sense of this word, as they vary exceedingly cross-linguistically. The biological constraints on such specific learning within the language domain — *universal grammar*, as this has come to be called, following Chomsky — are less relevant to Harris's research program, as I understand it. Indeed, along with Bloomfield who held that languages could vary arbitrarily, Harris may have doubted the reality of such constraints.

again as children ascend the linguistic hierarchy, acquiring the words, the phrases, and the sentence structures:[3] For example, by 9 months English-learning infants prefer to listen to sequences of unknown words from their own language rather than to Dutch words that violate the phonotactics of English (Jusczyk 1997). Infants, like adults, may identify word boundaries by recognizing words as familiar sound sequences. In a now justly famous experimental series, 8-month olds listened *for two or three minutes only* to an uninterrupted sequence of (computer generated) nonsense syllables constructed by randomly concatenating four nonsense words, e.g., *bidakupadotigolabubidaku* (Saffran, Aslin, & Newport 1996). The stimuli were presented with no pauses or variations in stress so as to offer no clue to word boundaries other than the high transitional probability linking syllables within but not across word boundaries. In test trials, infants heard words from this artificial language on some trials and 'part words' on other trials. Part words were sequences that had occurred during the familiarization phrase by chance concatenation of the four words that formed the basis of the training language. Infants showed a novelty effect in these experiments: They listened longer to the part-word test items showing that they had picked up on the distributional regularity after a very few exposures, with no aid from other phonological properties such as rhythmic structure or meaning — which for these stimuli didn't exist. In later replications, the frequency of word and word-part test items was controlled to insure that transitional probability, not simple frequency, was responsible for infant preferences. Such abilities to recover elements from their distributional patterns is not limited to language (similar effects are found for tone sequences) or to humans (tamerin monkeys behave much like human babies in these studies, Hauser, Newport, & Aslin 2001), but the domain- and species-generality of such skills and inclinations do not diminish their critical importance for initiating human language learning.

3. Actually, infants seem to accomplish this learning in a catch-as-catch-can procedure that may intermix discovery and organization of units at different levels of complexity over time, depending on how and when the evidence reveals itself, though overall the process is roughly bottom up. This sometime violation of strict bottom-up discovery is of course true also, as Harris writes, of working linguists, "they will usually know exactly where the boundaries of many morphemes are before they finally determine the phonemes." Nevertheless Harris saw a bottom-up procedure as one required to check the validity of obtained linguistic results. We don't know if infants are quite so methodologically demanding with their linguistic inductions.

Other experiments have shown that infants are also sensitive, by the same or related measures, to rhythmic properties and stress properties of the language (even where the syllables vary in segmental content, e.g. *KOgati . . . GAkoti*) and to global phonetic patterning (e.g., the tendency of the syllable sequences to follow an AAB versus ABA patterning, Marcus 2000). Analyzing each of these properties of the sound stream provides infants with some starting units that make possible ever more fine-grained analysis for words. The result is that babies well before they give evidence of knowing the meanings of words store previously unfamiliar words (*qua* phonetic sequences) heard in running speech and can recognize them when presented in isolation weeks later (Jusczyk and Aslin 1995).

The picture that emerges from the current literature is thus one of a perceptual learning process sensitive to interacting distributional regularities at many levels of analysis. Infants detect and make use of a variety of probabilistic cues that make some analyses of a speech sequence into words more likely than others. These cues include the internal consistency of phoneme or syllable sequences within familiar words, the typical stress patterns of native-language words, and probabilistic phonotactic regularities. I think that even Harris would have been surprised at these detailed confirmations of the scope — and psychological potency — of distributional analyses carried out by the infants of the species to acquire their native tongue.

1.2 Higher-order units

As soon as some words can be identified, children begin learning how these words are used both syntactically (how they are distributed with respect to other words and phrases) and semantically (how they function to convey meaning). These facts feed back into the word identification process allowing word selection to profit from knowledge of semantic, syntactic, and even discourse properties (which is how we can detect the speaker saying *grey tabby* versus *great abbey*, or *The sky is falling* versus *This guy is falling*, in dialects in which the phonetic renditions are highly overlapping.). One particularly fertile hypothesis (known as the 'prosodic bootstrapping hypothesis') again derives rather straightforwardly via certain extensions of Harrisian distributional procedures to implicate further units. The 'prosodic bootstrapping hypothesis' proposes that infants' perception of the increasingly familiar rhythmic and intonational structure of phrases and utterances in the exposure language guides analysis of the syntax (Gleitman, Gleitman, Landau, & Wanner 1988;

Morgan, Meier, & Newport 1987; Morgan 1986). The boundaries of utterances are well marked by salient acoustic cues. In English, words tend to be lengthened at the ends of phrases and at major phrase boundaries (Fisher & Tokura 1996). These prosodic boundaries help utterances cohere as perceptual units for adults, and even for 6- to 9-month olds. For example, Hirsh-Pasek et al. (1987) showed infant sensitivity to ½-second pauses in child-directed speech at utterance or phrase boundaries versus at arbitrary points within the sentence or phrase. Using the usual habituation tasks, 2-month olds were better able to remember the phonetic content of words produced as a single utterance than of the same words produced in a list or as two utterances (Mandel, Jusczyk, & Kemler-Nelson 1994); acoustic cues to the boundary locations of phrases are more variable and more subtle, but infants can perceive these as well (e.g., Gerken, Jusczyk, & Mandel 1994).

Although phonological similarity influences syntactic categorization, these sound-organizational properties are partial and probabilistic, and vary across languages. Semantic evidence for word classification is also probabilistic — for instance, why is *thunder* but not *lightning* a verb? Evidently children therefore continue to put most of their money on the more stable facts about relative distribution to establish the lexical classes that figure in phrase structural representations. The most familiar modern statement of a distributional learning algorithm for grammatical categories is from Maratsos & Chalkley (1980) whose schema directly follows linguists including Bloomfield (1933) and Harris (1951). They proposed that children could sort words into grammatical categories by noting their co-occurrences with other morphemes and their privileges of occurrences in sentences. Thus *–ed* is (probabilistically speaking) a verb-follower and *the* is a noun preceder. That these analyses are carried out by children even where supportive semantic evidence is absent — such as the notoriously semantics-free masculine-feminine distinction in many languages — makes a strong case for the continuing power of distributional analysis at higher linguistic levels. Children learn the grammatical gender classes at about the same rate that they pick up other grammatical categories. Evidently distributional analysis also operates several levels up from the phonetic distinctions with which it begins. Just as Harris proposed for linguistics-internal descriptive purposes, there is by now abundant evidence that probabilistic distributional analyses operate again and again to build up phonetic, syllabic, word, and sentence-like units and — by virtue of the same procedures — to grasp the distributional properties within each level.

2. Acquiring the meanings

After learners have parsed the sound wave of speech distributionally and prosodically so as to recover its linguistically functioning formatives, how do they assign interpretations to its elements and elementary structures, so discovered? How do they learn what words mean, and how these words function in the semantics of the clause? Clearly the primitive grounding for word meanings lies in the child's natural capacity to interpret scenes, properties, and events in the ambient extralinguistic world. However, there are many reasons to believe that this word-to-world learning procedure is too limited to carry the full load in explaining acquisition of the lexicon. The limitations of raw observation as the sole basis for lexical learning were driven home to me and my collaborator, Barbara Landau, when we discovered that the first verb in a blind child's vocabulary was liable to be *look* or *see*, and that these words — learned from the normal usage of sighted parents — were used sensibly to describe haptic exploration and apprehension (Landau & Gleitman 1985).

The robustness of lexical learning to differences in observational learning opportunities is only one of several reasons to suppose that, beyond the simple object-concept labels, acquisition at this level must be drawing on information beyond straightforward observation of scenes. For instance, how would perception of the ambient scene reveal the meaning of a verb like *know* or *think*? And there is a principled problem for acquiring perspective verbs such as *chase*/*flee*, *give*/*get*, *buy*/*sell* from observation alone for whenever the dogs are chasing the fox, the fox is fleeing the dogs. Should the fox turn and make a stand, the dogs can no longer be said to be chasing it. In short, these verb pairs which differ only in the perspective from which single scenes are viewed cannot be prised apart in terms of differing real-world contingencies. How are they acquired?

Distributional analysis is refined enough in principle to carry some of the learning burden here. On thinking about the blind learner's competence with words like *look* (and even *green* and *color!*), I was reminded straightaway of a parlor game that I developed, not coincidentally, while a member of Harris's lab group. We would render some structure in phrasal terminology, say

NP_i V NP_{ii} from NP_{iii}.

and the contestants were to think up some candidates for the verb, here, perhaps, *borrow* or *carry*. Next the range was to be narrowed by insisting that

the same N and V choices had to go also, either as paraphrase or preserving entailments, for some further frames:

NP_i V NP_{iii} from NP_{ii}

and

NP_i V NP_{ii} and NP_{iii}.

These would exclude *borrow* as a choice and instead bring to mind such items as *divide*, *distinguish*, and finally, with one or two more frames thrown in, maybe

NP_{ii} V from NP_{iii}

one converged on one or a very few verbs — aha, *separate!* Distributional overlap in complement structure has the effect of imposing a rough semantic classification not only on the major lexical classes, but proceeding on down to small subclasses: *Verbs of a feather flock together.*

In the following section I will describe one experimental series conducted in my lab that reveals something of the progress of lexical learning and which embodies among its most crucial properties successive distributional analyses of the kind prefigured in Harrisian thinking.

The outcome of this process is a probabilistic multiple-cue system (a constraint-satisfaction machinery; Gillette, Gleitman, Gleitman, & Lederer 1999; Snedeker & Gleitman, in press) for the interpretation of words. I have used the label *syntactic bootstrapping* both to describe the learning procedure that builds this cue system and for its efficient use, once constructed (probably in the child learner by age three years or so; Gleitman 1990). Restrictions and deficits in early child vocabulary learning are, on this view, attributable to incompletenesses in the cueing system, not — or not so much — to conceptual limitations in the learners.

2.1 Early noun dominance

The earliest child vocabulary (the first 50 to 100 words) is dominated by items that in the adult language we call nouns (Caselli et al. 1995; Gentner 1982). Noun proportion is at this stage strikingly higher than would be predicted from the input. Verbs are almost completely absent. The noun advantage in early speech and comprehension is robust across individuals and across languages, and is of great magnitude. All parties are agreed that

a major part of the explanation for this phenomenon will allude to the typical object-reference function of many frequently-encountered nouns. Perceiving the world in terms of concrete objects so as to relate them to their linguistic labels seems to be as natural as perceiving the acoustic world in terms of segmental properties that yield up the phonemes. Beyond this naturalness idea, explanation of the early noun-dominance effect in lexical acquisition splits into three broad approaches: The first is that the object concepts are developmentally primary, and so are represented adequately by even the youngest learners; the verbs, because they express relationships among these objects, are conceptually available only later in the child's mental growth. The second approach makes reference to the variable encoding of predicates across languages compared to the (near) identity of object-term reference across languages. Because the learner has first to discover the variable encoding properties for verbs in his own language, their acquisition is to that extent delayed. A third idea, which is my own and that of my collaborators, is that the informational state of the learning procedure limits the kinds of word that can be acquired early on. All of these approaches have something to recommend them, and they are not logically in opposition. It wouldn't be surprising if a conspiracy among them turned out to be the best explanatory theory. But here I will examine the third position: verbs are acquired later than nouns because their efficient learning demands sophisticated linguistic representations of the input. The required 'sophisticated' linguistic representations must themselves be constructed by using lower-level representations as the scaffolding (Gleitman 1990; Fisher, Hall, Rakowitz, & Gleitman 1994).

2.2 Efficient word learning

More or less coincident with the first rudimentary signs that children appreciate something of linguistic structure, somewhere around the second birthday, word learning becomes faster (increasing from about .3 words a day to 3 words a day), more precise, and more categorially catholic (nouns, verbs, adjectives all make their appearance). The temporal contiguity between structural appreciation and efficient, categorially broad, word learning has often been noted (e.g., Gleitman & Wanner 1982; Lenneberg 1967) and conjectured to reflect a cause-and-effect relationship. Our work seems to support this position, and is suggestive for how knowledge of word-cooccurrence and syntactic structure can inform vocabulary learning.

2.3 The Human Simulation paradigm

The findings I now review are from experimentation in which adult subjects try to identify words from partial information (Gillette, Gleitman, Gleitman, & Lederer 1999; Snedeker & Gleitman, in press). Conceptually, these experiments are analogous to computer simulations in which a device, endowed with whatever 'innate' ideas and learning procedures its makers deem desireable to program into it, is exposed to data of the kind naturally received by the target learner it is simulating. The measure of success of the simulation is how faithfully it reproduces the learning function for that target using these natural data. Our test device is a population of undergraduates (hence *Human* Simulations). Their preprogramming includes, *inter alia*, knowledge of English. The data received are contextualized mother-to-child speech events. The form in which the adult subjects receive information about these speech events is manipulated across conditions of the experiment. The subjects' task is to identify the mother's word meanings under these varying presentation conditions; that is, to 'acquire' a first vocabulary.

These experiments serve two purposes. The first is to provide an estimate of the psychological potency of various cues to word meaning that are latent in the real learning situation. For example, what kinds of words can — in principle — be acquired by inspection of the contingencies for their use in the absence of all other cues? The second is to estimate by reference to these outcomes something about the learning procedure used by children. Restating, we attempt to reproduce in adults the learning function of the one- and two-year old child by appropriate changes in the information structure of the input. If successful, this exercise makes plausible that the order of events in child vocabulary acquisition (here, the developmental move from nominal categories to predicate categories in speech and comprehension) is assignable to information-structure developments rather than to cognitive developments in the learner, for we have removed such possible cognitive inadequacies from the equation. When our college-age subjects fail to learn, it is not because they have some *conceptual* deficit that disbars them from recognizing words like *ball* or *get*. In fact, we know that these words are already in our subjects' vocabularies. All they have to do is to recover what they previously knew. Our first experimental probe asked how well they can do so by using only their ability to parse out relevant aspects of the passing scene.

2.4 Simulating word learning

The stimuli for these experiments were generated by videotaping mothers interacting with their 18 to 24-month old children in an unstructured situation. The maternal speech was transcribed to find the 24 most frequent nouns and the 24 most frequent verbs. To simulate a condition under which learners were presumed able only to identify recurrences of the same word in the speech stream, as in Saffran et al., and to match these with their extralinguistic contexts of use, we selected more or less at random 6 videoclips during which the mother was uttering each of these words. Each videoclip started about 30 seconds before the mother uttered the word, and ended about 10 seconds afterwards. That there were 6 clips for each 'mystery word' our subjects were to identify was to simulate the fact that real learners aren't forced to acquire meanings from a single encounter with a word and its context; rather, by examining the use of the word in a variety of contexts, the observer can attempt to parse out that property of the world common to all these encounters. The 6 video-clips for the word were then spliced together with a brief color-bar between them, and the subjects made a conjecture as to the word meaning after viewing all six samples; this procedure was repeated for all 48 items. The subjects were told for each word whether it would be a noun or a verb, but they *did not hear what the mother was saying, for we turned off the audio.* Subjects only heard a beep, which indicated the instant during the event when the mother actually had uttered the mystery word. So this manipulation should be thought of as simple word-to-world (rather: beep-to-world) pairing where the only information available to the learner is a sample of the extralinguistic contingencies for the utterance of words.

Of course, even if this stripped down situation fairly models an early stage in acquisition, namely, one in which the learner can't take advantage of the other words in the sentence and their syntactic arrangement (because she as yet knows neither), in the real case learners may receive 7 or 50 or 500 such word-to-world opportunities as the basis for inferring a word meaning. Our subjects received only 6. The only realistic question to ask of the outcomes, therefore, is not about the absolute level of learning, but only whether nouns and verbs are equally easy to identify under these presentation conditions.
The findings from this manipulation, based on 84 subjects observing the 48 items in 6 contexts each, are that about 45% of the nouns but only 15% of the

verbs were correctly identified. Moreover, each of the 24 nouns was identified by at least some of the subjects whereas eight — fully a third — verbs (*know*, *like*, *love*, *say*, *think*, *have*, *make*, *pop*) were never correctly identified by any subject. It is easy to see why: Several of these impossible verb*s* describe invisible mental acts and states while others are so general that they can be used in almost any context. If the only information available to youngest learners is inspection of these contexts of use, how could they be learned? As the results show further, even the 16 remaining verbs were considerably harder to identify from observation of their contexts (23% correct) than the nouns, which were identified correctly 45% of the time. So a noun bias in word identification can be demonstrated for any learner — even an adult — so long as real-world observation of that word's contingencies of use is the sole cue. If this is the early machinery for mapping between conceptual structure and word identity, we can understand the noun bias.

2.5 Imageability

There is considerable variability in identifiability within the noun class e.g., *ball* is correctly identified by just about every subject, but *kiss* by very few) and within the verb class (e.g., *push* vs *want*). This suggests that sheer concreteness or imageability rather than lexical category is the underlying predictor of identifiability of words from their observed contexts. And indeed a direct test shows this to be the case. We had a new group of subjects rate all the 48 items on a scale of imageability. Owing to the fact that the 48 test items were all highly frequent ones in speech to children under two years, not surprisingly the scale was heavily skewed to the imageable end. Nonetheless, almost every verb of the 24 in our set of commonly used verbs was rated as less imageable than almost any of the 24 nouns. As is obvious from this, an analysis for the predictive value of imageability and lexical class for identifiability showed that, once the imageability variable was removed, there was no effect of lexical class at all. That is, observation serves the learner well only for identifying the words that encode observeables! The correlation, in maternal speech, of this kind of observeability with the noun-verb categorization difference is so strong as to explain the predominance of nouns in early child speech, without having to consider whether — in addition — the children find the object terms less conceptually taxing.

2.6 Building representations that will support verb learning: the role of selection

An important outcome of the earliest stages of vocabulary learning is a well learned stock of concrete common nouns, and an unsystematic smattering of other items. But why doesn't the learner go on like this forever? The noun bias seems to be disfunctional for acquiring the verbs, the adjectives, and so forth. At least part of the answer is that the known nouns form the scaffold for building improved representations of the *linguistic* input; this more sophisticated input representation, in turn, supports the acquisition of more abstract vocabulary items. How could this progression work?

We know that, cross-linguistically, the melody and rhythm of maternal speech quite regularly delivers a unit of approximately clause size to the ear of the infant listener (Fisher & Tokura 1996). This unit, grounded in prosody, provides a domain within which learners can consider unknown verb meanings in the contexts provided by the nouns learned by the original word-to-world pairing procedure. Thus the acquired noun knowledge can ground a bootstrapping operation that carries the learner to a new representation of input speech by distributional analysis in the Harrisian sense.

One kind of information latent in such cross-word within-clause comparisons is *selectional:* Certain kinds of nouns tend to cluster with certain kinds of verbs within the clause. Notoriously, for example, *eat* is likely to occur with food words. A noun like *telephone* can be a give-away to a small class of common verbs, such as *talk*, *listen*, and *call*. The number of nouns in the maternal sentence (that is, within the prosodic contour of the clause) provides an additional clue to the verb meaning. This is because in the very short sentences used to infants, the number of nouns is a soft indicator of the number of arguments. Thus *gorp* in a prosodically bounded sequence such as . . . *John* . . . *gorp* . . . is more likely to mean "sneeze" than "kick". And "kick" is a better guess than "sneeze" for either . . . *John* . . . *gorp* . . . *ball* . . . or . . . *ball* . . . *gorp* . . . *John* . . . even if, because of insufficient language-specific syntactic knowledge, one cannot tell one of these last two from the other.

We next simulated a learner at this hypothetical stage of language acquisition: one who by prior acquisition of nouns and sensitivity to prosodic bounding information can register the *number* and the *meanings* of the nouns that characteristically occur in construction with the unknown verbs. To do so, we showed a new group of adult subjects the actual nouns that cooccurred with 'the mystery verb,' within the same six sentences for which the prior

subjects had seen the (videotaped) scenes. Within sentence, the nouns were alphabetized. The alphabetical order was chosen (as subjects were informed) because we were here modelling a device which has access to the semantics of noun-phrases that occur in construction with a verb, but has not yet acquired the language-specific phrase structure. Showing the actual serial orders would, for English adults, be tantamount to revealing the phrase structure. For example, the presentation form for the six maternal sentences containing the verb *call* was:

1. Grandma, you.
2. Daddy.
3. I, Markie.
4. Markie, phone, you.
5. Mark.
6. Mark.

Rather surprisingly, in light of the fact that these subjects saw no videotape, i.e., had access to no extralinguistic context at all, identifiability scores for verbs were as high in this condition as in the videotape condition. Much as Harris proposed based on corpora of adult-to-adult speech, there is information in adult-to-infant speech, in terms of the noun choices for a verb, even when the observer has no basis for identifying the structural positions in which these nouns occurred.

I hasten to acknowledge that there is no stage in language acquisition during which children hear nouns-in-sentences in alphabetic order and out of context. The manipulation was designed to extract — artificially, to be sure — a single information source (noun selectional information) and examine its potency for verb identification, absent other cues. Supposing that children make use of this information source, available as soon as they have acquired a modest stock of nouns, they can begin building a verb vocabulary.

2.7 Scene interpretation in the presence of noun knowledge

In the next manipulation, we modeled a learner who can coordinate the two sources of evidence so far discussed, inspecting the world to extract salient conjectures about relevant events, and using the known nouns to narrow the choice among them. New subjects were shown the 6 videotaped extralinguistic contexts along with the nouns the mother uttered in each sentence (in alphabetical order as before), and asked to identify the mystery verbs. Armed with

these dual information sources, subjects for the first time identified a respectable proportion (about 30%) of the verbs, a significant improvement over performance when either of these two information sources was made separately available. Noun knowledge evidently can serve as anchoring information to choose between plausible interpretations of the observed events.

2.8 Phrase structure representations support verb learning

The child as so far modeled has a principled basis for distinguishing between verbs that differ in argument number (e.g., *sleep*, *hit*, and *give*) and argument selection (e.g., *eat* occurs with food nouns, *call* occurs with *telephone*). This information together with inspection of scenes provides a principled basis for the construction of (certain aspects of) clause-level syntax. This is because a learner who

– understands such nouns as *ball* and *boy*
– hears an adult say *the boy hit the ball*
– observes some boy hitting some ball

can begin to locate the canonical subject of sentences in the language, that is, to label the phrases in a way that is relevant to thematic roles.

A correlated language-internal cue to subjecthood is that different categories of nouns probabilistically perform different thematic roles, a factor whose influence can be observed even in rapid on-line parsing performance among adults (Trueswell, Tannenhaus, & Garnsey 1994). Specifically, animate nouns are vastly more likely than inanimates to appear in subject position just because they are likely to be the causal agents in events (Dowty 1991). Once the position of the sentence subject is derived by matching up the observed agent with its known noun label, the young learner has a pretty good handle on the clause-level phrase structure of the exposure language.

The next experimental conditions tested the efficacy for verb identification of this phrase-structural information. The procedure directly follows the verb-guessing game that developed among Harrisian graduate students probing the power of distributional analysis at this level. First we showed a new group of subjects nonsense frames for our test verbs. These frames were again constructed from the sentences the mothers were uttering in the test conditions described earlier, by preserving the morphology but converting both the nouns and the verbs of the six maternal sentences to nonsense words. For example, two of the six stimuli for *call* were *Why don't ver gorp telfa?* and *Gorp wastorn,*

gorp wastorn! As Lewis Carroll — as well as Harris — would have predicted, there was a dramatic improvement under these conditions of verb identifiability, with subjects now identifying just over half the 24 verbs.

It may seem surprising that the syntactic environments of verbs should all by themselves be so informative of the verb meanings. Recall that in the manipulation we are describing, subjects saw no video — they knew nothing of the contexts in which these nonsense verbs had been uttered. And they knew none of the coocccuring nouns, for all these had been converted to nonsense too. Yet the subjects identified proportionally more verbs (51%) than they did in the prior experiment, in which both video contexts and accompanying nouns were presented (29%).

On closer consideration, it makes sense that syntactic information can provide major cues to verb meaning. Verbs differ in the structures in which they appear. Generally speaking, the closer any two verbs are in meaning, the more their structural privileges overlap (Fisher, Gleitman, & Gleitman 1991). This is because the structural privileges of a verb (the number, type, and positioning of its associated phrases) derive, quirks and provisos aside, from an important aspect of its semantics; namely, its argument-taking properties. The *number of argument positions* lines up with the number of participants implied by the logic of the verb. Thus a verb that describes a self-caused act of the musculature (e.g., Joe *snoring*) is liable to surface intransitively, a physical effect of one entity on another (Joe *throwing* a ball) is likely to be labelled by a transitive verb, and an act of transfer of an entity between two places or persons is likely to be ditransitive (Joe *giving* a ball to Bill). The *type of complement* is also derivative of aspects of the verb's meaning. Thus a verb describing a relation between an actor and a proposition is likely to take clause-like complements (Joe *believing* that Bill is sad). And the hierarchical arrangements of the noun phrases cues the thematic role of the participant entities. Because verb meanings are compositional at least at the level of these argument-taking properties (Grimshaw 1990), the matrix of verb-to-structure privileges has the effect of providing a coarse semantic partitioning of the verb set. For example, because one can forget *things*, this verb licenses a noun-phrase complement; and because one can also forget *events*, the same verb also licenses clausal complements.

A vast linguistic literature documents these syntax-semantics relations (see, e.g., Gruber 1967 and Fillmore 1968 for seminal discussions; Croft 1991, Goldberg 1995, and Levin 1993 for recent treatments; for experimental documentation, Fisher et al. 1991; and for cross-linguistic evidence concerning caretaker speech, Lederer, Gleitman, & Gleitman 1995; Geyer 1996; Li 1994;

for learning effects in young children, Brown 1973; Fisher, Hall, Rakowitz, & Gleitman 1994; Naigles 1990; Waxman 1999). In the simulation, subjects were able to use syntax and associated morphology to make inferences about the verb meanings even though they were artificially prohibited from observing the contexts of use and the cooccurring nouns.

2.9 Coordination of cues in efficient verb learning

Two further conditions fill out the set of simulations. Our next pool of adult subjects saw the full sentences that the mothers had been uttering, i.e., we popped the real nouns back into the frames of the previous condition, leaving only the verb as nonsense. From this information, and without video context, subjects correctly identified three quarters of the verbs. Finally, we provided a last group of subjects with the information that we believe is available to the two and three-year old learners: full linguistic representations of the sentences along with their (videotaped) extralinguistic contexts. Now the mystery verbs were no mystery at all, and the subjects succeeded in identifying about 90% of them.

3. The distribution of semantic information in language design: Why the imageable nouns are acquired first

Taken together, the experiments just presented model a learning device that is seeking word meanings by convergence across a mosaic of probabilistic linguistic-distributional and situational evidentiary sources, a constraint-satisfaction device. However, the learning device is not supplied by nature with the ability to use all these sources of information from the beginning. Rather, it has to construct the requisite representations on the fly during the process of learning, for languages differ not only in how they pronounce the words but in how they instantiate predicate–argument structure in the syntax of sentences.

One source of information about word meaning is in place from the beginning of the word learning process and constitutes the ground on which the learner stands to build the full lexicon: This is the ability to interpret the world of scenes, objects, properties, and events, and to suppose that these will map regularly, even if complexly, onto linguistic categories and structures. This initial information source serves two interlocking purposes. First, it allows the learner to acquire a very special kind of lexical-semantic information, that which can be gleaned by pragmatically informed perception: names

for object concepts. Second, the information so acquired enters into an incremental process for building the clause-level phrase structure of the language. The structure building processes are based on the supposition that form-to-meaning relations will be as transparent as possible; for example that the relations between the number of participants in the scene and the number of noun phrases in the clause will be one-to-one, in the best case. These ideas are maximally stripped down renditions of principles that in linguistics go under such names as the projection principle and the theta criterion (Chomsky 1981). Moreover, form-to-meaning transparency also requires, in the best case, that conceptual dominance relations among actors in an event (the agent, the theme or patient, the recipient) map onto structural dominance in the sentence (the subject, the direct object, the indirect object); a thematic hierarchy.

Each of the data sources from which this knowledge is built is errorful, of course. People sometimes talk of things, even concrete things, when these are absent. The number of nouns in a sentence, even a short sentence, isn't always a straightforward reflection of argument number. The syntactic mapping of verb-argument structure onto surface structure is quite quirky, e.g., in English the verb *butter* surfaces with one too few arguments and *rain* with one too many.[4] Yet, as shown for adults in the simulations just described, coordinated use of these several cues rapidly converges to verb identifiability. We suspect that the same is true for child learners.

More generally, the burden of the experiments was to show that the lexicon-building task is naturally partitioned into a relatively concrete subpart, which requires *little linguistic knowledge* and thus is acquired first, and a more abstract part which is acquired later because its discovery requires support from *linguistically sophisticated representations*. In these experiments, words

4. The predicate *rain* is argumentless (no body, no god, is its agent or experiencer), but the requirement of English for an overt subject noun-phrase necessitates the dummy ('expletive') *it* for *rain* and other weather terms; hence, sentences containing such verbs as *rain* and *snow* contain one more argument than their logic requires. As for *butter*, consider related items such as *cover* or *spread* (He covered/spread his bread with butter). The denominal verb *butter*, despite the fact that like these other instances it logically requires three arguments, occurs with one fewer because the third argument is incorporated into the morphology of the verb. (We don't have to say *I buttered my bread with butter*; under most — though not all — ordinary circumstances, *I buttered my bread* would usually carry the implication *with butter.*)

like *want*, *know*, and *see* were not merely 'facilitated' compared to words like *push* and *go*, when phrase-structural information was made available. The abstract verbs (e.g., *think* or *want*) that were literally *impossible* to learn from observation (zero percent correct in the video-beep condition) were significantly the *easiest* to learn (identified close to 100% of the time), compared to the concrete verbs (e.g., *go* or *push*), in all three conditions in which the subjects had syntactic information.

I imagine that if Harris saw these findings (and if he could for a moment stifle a natural modesty) he would have responded "Obviously." And he would have been literally correct had he said "I told you so." Yes he did, clearly and in detail.

References

Aslin, R.N., J.R. Saffran, & E.L. Newport. 1998. "Computation of conditional probability statistics by 8-month-old infants." *Psychological Science* 9:321–324.

Baillargeon, R. 1998. "Infants' understanding of the physical world". In M. Sabourin, F. Craik, & M. Robert (Eds.), *Advances in psychological science*, Vol. 2, pp. 503–529. London: Psychology Press.

Bloomfield, L. 1933. *Language*. New York: Holt.

Brown, R. 1973. *A first language*. Cambridge, MA: Harvard University Press.

Casselli, M.C., E. Bates, P. Casadio, J. Fenson, L. Fenson, L. Sanderl, & J. Weir, 1995. "A cross-linguistic study of early lexical development". *Cognitive Development* 10: 159–199.

Chomsky, N. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Croft, W. 1990. *Typology and universals*. New York: Cambridge University Press.

Dowty, D.1991. "Thematic proto-roles and argument selection". *Language* 67:547–619.

Eimas, P.D. 1975. "Auditory and phonetic coding of the cues for speech: Discrimination of the [r–l] distinction by young infants". *Perception and Psychophysics* 18:341–347.

Feldman, H., S. Goldin-Meadow, & L.R. Gleitman. 1978. "Beyond Herodotus: The creation of language by linguistically deprived deaf children". In A. Lock (Ed.), *Action, symbol, and gesture: The emergence of language*, pp. 351–414. New York: Academic Press.

Fillmore, C.J. 1968. "The case for case". In E. Bach & R.T. Harms (Ed.), *Universals in linguistic theory*, pp. 1–88. New York: Holt, Rinehart & Winston, Inc.

Fisher, C., H. Gleitman, & L.R. Gleitman. 1991. "On the semantic content of subcategorization frames". *Cognitive Psychology* 23:331–392.

Fisher, C., D.G. Hall, S. Rakowitz, & L.R. Gleitman. 1994. "When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth". *Lingua* 92:333–375.

Fisher, C., & H. Tokura. 1996. "Acoustic cues to grammatical structure in infant-directed speech: Cross-linguistic evidence". *Child Development* 67:3192–3218.

Gentner, D. 1982. "Why nouns are learned before verbs: Linguistic relativity versus natural partitioning". In K. Bean (Ed.), *Language, Thought, and Culture*, pp. 301–334. Hillsdale, NJ: Erlbaum.

Geyer, H. 1996. "Subcategorization as a predictor of verb meaning". Unpublished ms., Univ. of Pennsylvania.

Gerken, L., P.W. Jusczyk, & D.R. Mandel. 1994. "When prosody fails to cue syntactic structure: 9-month-olds" sensitivity to phonological versus syntactic phrases". *Cognition* 51: 237–265.

Gillette, J., H. Gleitman, L.R. Gleitman, & A. Lederer. 1999. "Human simulations of vocabulary learning". *Cognition* 73: 135–176.

Gleitman, L.R. 1990. "The structural sources of verb meanings". *Language Acquisition* 1: 3–55.

Gleitman, L.R., & H. Gleitman, 1997. "What is a language made out of?" *Lingua* 100: 29–55.

Gleitman, L.R., & E. Wanner. 1982. "Language acquisition: The state of the state of the art". In E. Wanner & L.R. Gleitman (Ed.), *Language acquisition: State of the art*, pp. 3–48. New York: Cambridge University Press.

Gleitman, L.R., H. Gleitman, B. Landau, & E. Wanner. 1988. "Where learning begins: Initial representations for language learning". In F.J. Newmeyer (Ed.), *Linguistics: The Cambridge survey, Vol. 3. Language: Psychological and biological aspects*, pp. 150–193. New York: Cambridge University Press.

Goldberg, A.E. 1995. *Constructions: A construction grammar approach to argument structure.* Chicago: University of Chicago Press.

Goldin-Meadow, S., & C. Mylander. 1984. *Gestural communication in deaf children: The effects and noneffects of parental input on early language development.* Monographs of the Society for Research in Child Development 49.

Grimshaw, J. 1990. *Argument structure.* Cambridge, MA: MIT Press.

Gruber, J.S. 1967. Look and see. *Language* 43: 937–947.

Harris, Z. 1951. *Methods in structural linguistics.* Chicago: University of Chicago Press.

Hauser, M.D., E.L. Newport, & R.N. Aslin. 2001. "Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins". *Cognition* 78. B53–B64.

Hirsh-Pasek, K., D.G. Kemler Nelson, P.W. Jusczyk, K. Wright-Cassidy, B. Druss, & L. Kennedy. 1987. "Clauses are perceptual units for young infants". *Cognition* 26: 269–286.

Jackendoff, R. 1983. *Semantics and cognition.* Cambridge: MIT Press.

Joshi, A. 1999. "A parser from antiquity", in A. Kornai (ed.) *Extended finite state models of language*, pp. 6–15. New York: Cambridge University Press.

Jusczyk, P.W. 1997. "The discovery of spoken language". Cambridge MA: MIT Press.

Jusczyk, P.W., & R.N. Aslin. 1995. "Infants' detection of the sound patterns of words in fluent speech". *Cognitive Psychology* 29: 1–23.

Kuhl, P.K. 1994. "Learning and representation in speech and language". *Current Opinion in Neurobiology* 4: 812–822.

Landau, B. & L.R. Gleitman. 1985. *Language and experience: Evidence from the blind child.* Cambridge, MA: Harvard University Press.

Lederer, A., H. Gleitman, & L.R. Gleitman. 1995. "Semantic information in the structure of maternal speech: Verbs of a feather". In M. Tomasello & W.E. Merriman, *Beyond words for things: Young children's acquisition of verbs*, pp. 277–297. Hillsdale NJ: Erlbaum.

Lenneberg, E.F. 1967. *Biological foundations of language.* New York: Wiley

Levin, B. & M. Rapaport Hovav. 1995. *Unaccusativity: at the syntax-lexical semantics interface.* Cambridge MA: MIT Press.

Li, P. 1994. *Subcategorization as a predictor of verb meaning: Cross-language study in Mandarin.* Unpublished manuscript, Univ. of Pennsylvania.

Mandel, D.R., P.W. Jusczyk, & D.G. Kemler-Nelson. 1994. "Does sentential prosody help children organize and remember speech information?" *Cognition* 53:155–180.

Maratsos, M. & M.A. Chalkley. 1980. "The internal language of children's syntax". In K. Nelson (Ed.), *Children's language*, pp. 1–28. New York: Gardner Press.

Marcus, G.F. 2000. "Pabiku and Ga Ti Ga: Two mechanisms infants use to learn about the world". *Current Directions in Psychological Science* 9:145–147.

Morgan, J.L. 1986. *From simple input to complex grammar.* Cambridge, MA: MIT Press.

Morgan, J.L., R.P. Meier, & E.L. Newport. 1987. "Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases in the acquisition of language". *Cognitive Psychology* 19:498–550.

Naigles, L.G. 1990. "Children use syntax to learn verb meanings". *Journal of Child Language* 17:357–374.

Newport, E.L. 1990. "Maturational constraints on language learning". *Cognitive Science* 14:11–28.

Saffran, J.R., R.N. Aslin, & E.L. Newport. 1996. "Statistical learning by 8-month-old infants". *Science* 274:1926–1928.

Senghas, A. & M. Coppola. (in press). "Children creating language: How Nicaraguan Sign Language acquired a spatial grammar". *Psychological Science* 12.4:323–328.

Senghas, A., M. Coppola, E.L. Newport, & T. Supalla. 1997. "Argument structure in Nicaraguan Sign Language: The emergence of grammatical devices". In E. Hughes, M. Hughes, & A. Greenhill (Ed.), *Proceedings of the Boston University Conference on Language Development*, pp. 550–561. Boston: Cascadilla Press.

Snedeker, J., & L.R. Gleitman. (in press). "Why it is hard to label our concepts". To appear in D.G. Hall & S.R. Waxman (eds.), *Weaving a lexicon.* Cambridge, MA: MIT Press.

Spelke, E.S., K. Breinlinger, J. Macomber, & K. Jacobson. 1992. "Origins of knowledge". *Psychological Review* 99:605–632.

Trueswell, J.C., M.K. Tanenhaus, & S.M. Garnsey. 1994. "Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution." *Journal of Memory and Language* 33:285–318.

Waxman, S.R. 1999. "Specifying the scope of 13-month-olds' expectations for novel words". *Cognition* 70: B35–B50.

Werker, J.F., & R.C. Tees. 1984. "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life". *Infant Behavior and Development* 7:49–63.

Part IV

Phonology

Chapter 12

The voiceless unaspirated stops of English*

Leigh Lisker
Haskins Laboratories

1. Introduction

Linguists are generally agreed that the English stop consonants fall into two phonological sets: /bdg/ and /ptk/. These two sets embrace a range of phonetic stop types, perhaps the largest number proposed being in Trager & Smith 1951:30–34). Thus for /b/ they find three allophones [b] [b̥b] [b b̥], while for /p/ they find four: [pʰ] [p] [**P**] [p˺]. [b̥b] and [bb̥] are marked by voiceless onsets and offsets respectively, though otherwise voiced, only [b] standing for a fully voiced stop, that is, one in which the interval of labial closure is largely accompanied by glottal signal, as per the definition of the International Phonetic Association; and all three are described as lenis (i.e. weakly articulated) and unaspirated. Three of the phonetic variants of /p/ are described thus: [pʰ] is voiceless, more or less aspirated and 'quite fortis'; [p] is voiceless fortis and unaspirated; and [p˺] is voiceless, fortis and not audibly released. Within words [p] occurs mainly in medial position before an unstressed vowel, and [p˺] is found most often prepausally. When a word terminating in /p/ immediately precedes a word with an initial vowel, /p/ will be represented by [p], sometimes with attendant glottalization at the onset of the vowel, though such glottalization has nothing to do with the stop, reflecting rather a speaker's intent to indicate the presence of a word boundary. The phonetic as well as the phonological status of [**P**], which is found only in clusters with a preceding [s] is said to be problematical, in that it is judged to be similar to [b] [b̥b] [bb̥] well as to [p], but assigned either by convention or on certain distributional grounds to

* The work reported in this chapter was supported by the National Institute of Health NIDCD under grant DC-02717 to the Haskins Laboratories. I want also to thank Leonard Katz for his patient counseling on matters statistical.

/p/ rather than /b/.[1] Nowadays many linguists and phoneticians represent both the [ʰb] and [bʰ] forms of /b/ by [b̥], while [**P**] is represented simply as [p]. For some observers initial [b̥] is 'voiced as in English', or 'voiceless and lax', or 'voiced but not pre-voiced'.[2] For such observers, then, the language-universal definition of 'voiced stop' long established by the International Phonetic Association gives way to a language-specific one. (See Keating 1984: 288.) To be sure, from articulatory studies evidence has been reported that the state of the glottis during the closures of initial voiceless allophones of /bdg/ may be much the same as it is for the truly voiced allophones of those phonological elements (Flege 1982), although from a strictly acoustic point of view they may be no more voiced than are [ptk]. Aside from any claim that [b̥d̥g̊] (phonologically /bdg/) are different phonetically from [ptk] (phonologically /ptk/), another, and for the linguist perhaps more compelling reason for not representing voiceless /bdg/ as [ptk] is purely phonological, in that this allows us to maintain that the aspiration occurring at the release of voiceless fortis stops in certain positions is 'automatic', i.e. provided by a phonotactic rule, and is hence not an independent segment (Bloch & Trager 1942; Harris 1951). Thus, although voiceless variants of /bdg/ may satisfy the IPA definitions of [ptk], we shall for now follow established custom in writing them phonetically as [b̥d̥g̊]. Moreover, since we do not at present have much evidence as to whether the voiceless unaspirated labial stop after /s/ constitutes a third type of voiceless unaspirated stop to be distinguished from either [b̥] or [p] or from both, we shall follow Trager & Smith 1951 in spelling it as [**P**]. The question to be addressed here is whether these three putatively different labial stops [b̥] [**P**] [p], all of which merit acoustically the description 'voiceless unaspirated stop', are perceptually (and perhaps otherwise?) distinguishable, when by editing they are all presented in utterance-initial position.

The following three sentences which include [b̥] [**P**] [p] were chosen because they contain the same number of syllables and all terminate in a rising intonation:

---

1. Several students of English, to be sure, assign the post-/s/ stops to the /bdg/ category on grounds of phonetic similarity, and find little merit in any distributional argument for /ptk/ (Hultzén 1962; Davidsen-Nielsen 1969).

2. However, no one describes the /p/ in such a word as *rapid* as being 'voiced but not prevoiced', although from an acoustic point of view it may be no less voiced than the initial stop of an initial *bid*.

(1) Did he win this bout?
(2) Did you fix the spout?
(3) Didn't you drop out?

Let us try to determine whether the intervals following their voiceless labial closure intervals, when presented in isolation, are perceptually distinguishable. An initial finding is that these post-labial closure intervals, when presented out of their original contexts, are all heard by English-speaking listeners to begin with a 'b'; i.e. they are identified as the word *bout*, and not *pout* or some other, so that per Harris's pair test, all three labial stop types (at least so far as their post-closure intervals are concerned) might be considered phonologically identical, although of course in their original contexts they are clearly different phonologically. The identification of the post-closure intervals as the word *bout* is, in the case of the first and second sentences, not at all surprising; (1) in the first sentence the final intended word was, after all, *bout*, and (2) it has long been known that when a word beginning with the cluster [s**P**] is deprived by waveform editing of its initial sibilant noise English-speaking listeners identify the residue as a word beginning with a /b/ (cf. Lotz, Abramson, Gerstman, Ingemann, & Nemser 1960; Davidsen-Nielsen 1969; Reeds & Wang 1961). That the post-labial closure interval of the third sentence should also be heard as *bout is* somewhat unexpected. The stop in its original context is clearly heard to be a member of /p/, and there is nothing in the phonetic literature (so far as I know) to suggest that the perceptual character of this stop, given that it is accompanied by a release into a following vowel, will nevertheless derive entirely from its closing transition and following silent closure interval, and that an opening transition might provide contradictory evidence as to the voicing state of the consonant. (There is, in fact, evidence that when the closure interval of an intervocalic sequence of two stops differing in place of articulation is reduced in duration, it is the opening transition of the second stop that prevails perceptually (Repp 1978).) On the other hand, given the context in which this transition is presented, it is not entirely surprising that we do not hear a 'p'. First of all, the absence of closure voicing, which in the original sentence is no doubt taken to be the result of a laryngeal devoicing gesture, can now be understood simply as part of the silence preceding speech onset. The now-initial stop release and transition are then heard as /b/, presumably because there is no aspiration. But one might on the other hand suppose that an initial voiceless unaspirated stop preceding a stressed vowel, particularly when the stop is described as fortis, need not necessarily be heard as 'b', for its burst and

transition might well be characterized by the intensity and F0 and F1 features said to mark voiceless unaspirated as well as aspirated stops (Fischer-Jørgensen 1968; Erickson 1975; Ohde 1984). For a demonstrated ability to distinguish among [b̥][p][**P**] one or another of the following explanations might be advanced: (1) that the /b/ of the first sentence is somewhat voiced (but not 'pre-voiced'), while the labial stops in the second and third sentences are entirely voiceless; (2) that the labial in the first sentence is lenis, while that of the third, and perhaps the second, are fortis; (3) that the /b/ of the first sentence is marked by a release burst intensity and a post-closure fundamental frequency contour different from that of the [p] in the third, and perhaps the [**P**] of the second as well. (To be sure, some linguists judge the stops found after word-initial /s/ to be voiceless and lenis — e.g Hultzén 1962; Schane 1968.)

Of course, we must bear in mind that in a good many other languages, e.g. Spanish, Polish, Dutch, Thai, Korean, there are initial stops that are regularly described as being voiceless and unaspirated, some of which English-speaking listeners identify sometimes with the English 'voiced' stop phonemes, and sometimes with the phonologically voiceless ones. Thus teachers of Spanish provide anecdotal evidence that some of their English-speaking students sometimes identify the voiceless stops in that language with the English 'voiced' stops. Indeed, a carefully controlled experiment in which the initial voiced and voiceless unaspirated stops of standard Dutch were labeled by a group of phonetically naïve English speakers revealed that the Dutch voiceless unaspirated stops were identified almost entirely as 'ptk' (Lisker 1979). Moreover, some exploratory experiments indicate that the initial voiceless unaspirated stops of Polish and Russian, which like the Dutch voiceless stops contrast with fully voiced (i.e.'pre-voiced") categories, are often identified as 'ptk' by English-speaking listeners, even when they are presented in the absence of the contrasting voiced stops. All this suggests that a cross-language application of the term 'voiceless unaspirated' is a phonetically inadequate characterization of voiceless stops with voice onset time values lying within the 0 to about +40 msec range, or even more in the case of velars.

2. Experiment

Ten tokens of each of the English sentences listed above were recorded over a period of several weeks in randomized orders. The speaker was a native speaker of American English, born and raised in a large urban center along the mid-

Atlantic coast. From the recorded target sentences the intervals following the labial closures were extracted and presented to a number of English-speaking listeners. As was stated above, these stimuli were heard as phonologically identical, all being identified as the word *bout*, so that, whatever the phonetic differences that might underlie the phonological differences among the labial stops [b̥] [p] [**P**] of the three sentences, those embodied in the releases and opening transitions were not robust enough perceptually to survive the extraction process. This finding does not allow us to say whether or not the post-closure stops are phonetically discriminable. To answer this question three tests were conducted — in one the stimuli presented derived from *Did he win this bout?* and *Did you fix the spout?*, in another they came from *Did he win this bout?* and *Didn't you drop out?*, and in the third they came from *Did you fix the spout?* and *Didn't you drop out?* In each test ten English-speaking listeners were given the task of deciding from which of the two source sentences each stimulus had been derived. In each test the response data, when subjected to an analysis of variance (ANOVA), indicated that the ten listeners as a group were unable to identify the test stimuli as to their source sentences. In other words, the main effect of Source Sentence was not significant — ($F(1,18) < 3.0$; $p > .10$). Thus the available perceptual evidence fails to indicate the presence of any phonetic feature or features by which English-speaking listeners can distinguish the releases and following transitions of the stops [b̥] [p] [**P**]. Consistent with this result is the finding that acoustical measurements of the test stimuli reveal no statistically significant differences in the timing of voice onset ('VOT', cf. Lisker & Abramson 1964) following the releases of the [b̥], [p], and [**P**] stops (by ANOVA $F(2, 27) = 1.93$, $p = .165$). To be sure, when the silent closures of these stops were measured in their original contexts, their durations were found to differ significantly, with mean values of about 70, 80, and 100 ms. for [p], [**P**], and [b̥] respectively ($F(2, 27) = 18.31$; $p < .0001$).[3] Whether the significantly greater labial closure durations in the first sentence play an important role in listeners' perception of [b̥] (= /b/) as against [p] (= /p/) of the second

3. VOT and closure durations were measured in productions of a second English speaker: no significant VOT differences among [p] [**P**] [b̥] were found ($F(2, 27) = .146$, $p = .865$), while closure durations for the three stops were 66 ms, 69 ms, and 79 ms respectively, the value for [b̥] being significantly greater than either of the others ($F(2, 27) = 3.43$, $p = .009$). Her /s/ durations were also significantly greater in Sentence 2 than in Sentence 1 (by ANOVA $F(1, 18) = 7.06$; $p = .016$).

sentence or [**P**] (phonological status moot) of the third sentence, is quite uncertain, but cannot be ruled out. Quite possibly, too, the significantly longer /s/ durations before the labial closures in Sentence 2 as compared to those in Sentence 1 (by ANOVA $F(1,18) = 80.40$; $p < .0001$) may also contribute to listeners' ability to distinguish between their two final noun phrases: *this bout* vs *the spout*.[4] To be sure, informal tests in which closure durations were manipulated had nil effect on sentence identification, while manipulations of the /s/ durations in Sentences 1 and 2 had only marginal effects on identifications of the two final words of those sentences.

3. Discussion

The finding that the releases and opening transitions (together with the following syllable codas) of [b̥] [**P**] [p], when presented as isolated stimuli, were not distinguished by English-speaking listeners, does not justify a conclusion that these stops, all of them acoustically voiceless and unaspirated, are therefore phonetically identical. (The literature might lead us to expect them to differ in the properties of their burst releases, the fundamental frequency contours of the post-release voicing, and also in their first-formant transitions, though acoustic measurements of our stimuli provide no evidence of such.) The failure of listeners to distinguish among the stimuli with [b̥] [**P**] [p] onsets, and the apparent absence of any readily measured acoustic differences, does not preclude the possibility that they differ in articulation, in that whereas the silence of the [b̥] closure of Sentence 1 can be understood to result from a devoicing gesture associated with /s/, the closing transition from the preceding vowel to the [p] closure of Sentence 3 must incorporate acoustic properties that signal a devoicing gesture which can only be attributed to the /p/ stop itself. As for the voiceless closure of [**P**], there is evidence that this, like that of [b̥], results from a devoicing maneuver associated largely, perhaps entirely, with the preceding /s/.[5] In fact, the glottal devoicing gestures for a

4. Quite possibly a difference in vowel quality between the unstressed vowels of *this* and *the* may also make a perceptual contribution.

5. To be sure, a once prevalent and possibly still current view is that the [**P**]'s voicelessness reflects an underlying /p/, and that the absence of aspiration is contextually determined by the preceding syllable-initial /s/.

word-initial prevocalic [s], an initial [s**P**], and a sequence of [s] and [b̥] across a word boundary may well be either phonetically identical or only insignificantly different (Yoshioka, Løfqvist, & Hirose 1981).

4. Summary

With respect to three English voiceless unaspirated labial stops that we have represented as [b̥][p] [**P**], when they, or rather their post-closure releases plus following codas, are presented in isolation, they are all perceived as instances of the phonological category /b/ by native speakers of English. Moreover, despite their differences in both phonetic and phonological status within the source sentences, they appear to be perceptually indistinguishable when presented in test stimuli where they are perceived as utterance-initial. In addition, these same intervals are in all three sentence contexts interchangeable without phonological effect. From these findings we may conclude that, whatever the phonetic differences commonly attributed to [b̥] [p] [**P**] in their differing contexts, deletion of the environments in which they were produced has removed differentiating features essential to the identification of the terminal portion of Sentence 1 as *bout*, that of Sentence 2 as *spout*, and that of Sentence 3 as *drop out*.

References

Bloch, Bernard & George L. Trager. 1942. *Outline of Linguistic Analysis*. Linguistic Society of America. Waverly Press, Baltimore, MD.

Davidsen-Nielsen, Niel. 1969. "English stops after /s/". *English Studies* 50:321–339.

Erickson, Donna. 1975. "Phonetic implications for an historical account of tonogenesis in Thai". In *Studies in Tai Linguistics in Honor of William J. Gedney* (Eds. J.H. Harris & J.R. Chamberlain, pp. 100–111. Bangkok: Central Institute of English Language Office of State Universities.)

Flege, James E. 1982. "Laryngeal timing and phonation onset in utterance-initial stops". *Journal of Phonetics* 10:177–192.

Harris, Zellig S. 1951. *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

Hultzén, Lee S. 1962. "Voiceless lenis stops in prevocalic clusters". *Word* 18:307–312.

Fischer-Jørgensen, Eli. 1968. "Les occlusives françaises et danoises d'un sujet bilingue". *Word* 24:112–153.

Keating, Patricia A. 1984. "Phonetic and phonological representation of stop consonant voicing". *Language* 60:286–319.

Lisker, Leigh. 1979. "Speech across a linguistic boundary: Category naming and phonetic description". In *Amsterdam Studies in the Theory and History of Linguistic Science* IV, Current Issues in Linguistic Theory, Vol. 9, (Eds. Harry & Patricia Hollien), 565–571.

Lisker, Leigh & Arthur S. Abramson. 1964. "A cross-language study of voicing in initial stops: Acoustical measurements". *Word* 20: 384–422.

Lotz, John, Arthur S. Abramson, Louis J. Gerstman, Frances Ingemann, & William J. Nemser. 1960. "The perception of English stops by speakers of English, Spanish, Hungarian and Thai: A tape-cutting experiment". *Language and Speech* 3: 71–77.

Ohde, Ralph N. 1984. "Fundamental frequency as an acoustic correlate of stop consonant voicing. *Journal of the Acoustic Society of America* 75: 224–230.

Reeds, J. A. & William S.-W. Wang. 1961. "The perception of stops after s". *Phonetica* 6: 78–81.

Repp, Bruno H. 1978 "Perceptual integration and differentiation of spectral cues for intervocalic stop consonants". *Perception and Psychophysics*, 24: 471–485.

Schane, Sanford A. 1968. "On the non-uniqueness of phonological representations". *Language* 44: 709–716.

Trager, George L. & Henry L. Smith, Jr. 1951. *An Outline of English Structure.* (Studies in Linguistics: Occasional Papers, 3). Norman, Oklahoma: Battenburg Press.

Yoshioka, Hirohide, Anders Löfqvist, & Hajime Hirose. 1981. "Laryngeal adjustments in the production of consonant clusters and geminates in American English". *Journal of the Acoustic Society of America* 70: 1615–1623.

Chapter 13

On the bipartite distribution of phonemes

Frank Harary and Stephen Helmreich
New Mexico State University

To the memory of Zellig Harris, an excellent mathematical linguist

1. Introduction

In December 1989, one of us was flying on Korean Airlines from Tokyo to Los Angeles and had no choice but to land in Honolulu because that was the only flight we could get during the holiday season. Since Honolulu is in the USA, we were obliged to go through customs there. This meant that all the baggage had to be taken from the plane into the airport. Each person had to find his checked luggage and go through customs with it. Then it was all reloaded onto the plane and some three hours later we took off again for LA. The process began with all of us being shepherded into several large buses where we heard the following recorded message:

> Aloha. Welcome to Honolulu, Hawaii, USA. You will be going through customs and the whole process will take about three hours. Please do visit our shops at the airport and get a souvenir to bring back to the mainland. Well again, welcome to Hawaii. Ahola.

This unexpected metathesis re-awakened our interest in the Hawaiian language, leading to our first essay on the present topic (1991), which we have now extended.

We have two objectives. First, we look at the consonant/vowel distinction from the standpoint of graph theoretic concepts. Building on the work of Harary & Paper (1957), we relate this distinction to the graph-theoretic

We are grateful to David Farwell and Bruce Nevin for their helpful suggestions.

notion of a bipartite graph. In a bipartite graph G every node can be colored with one of two colors so that each edge of G joins two nodes of different colors. In this context the colors are vowel and consonant. We study the graph in which the nodes are phonemes and the edges are determined by succession in the corpus. We introduce a method of quantifying the degree of bipartiteness of the phonemic graph of a particular corpus. We apply this method to Hawaiian. We show that the consonant/vowel division produces a highly bipartite graph.

Second, we generalize this result by developing a program which examines divisions of elements of a set into two groups, in this case, phonemes, including such phonetic divisions as front/back, high/low, etc. It determines the bipartiteness of each division with respect to the graph of a corpus in that language. We show that the consonant-vowel division produces the most bipartite graph for Hawaiian and examine several other languages as well. This approach provides a distributional method of identifying this distinction.

Section 2 gives the linguistic and graph theoretic backgrounds. Section 3 describes the methodology with reference to our initial work on Hawaiian, and the extension of that approach to a more general thesis, namely, that in any natural language, L, a partition of the phonemes of L into consonants and vowels is the most bipartite division. Section 4 presents results for several languages. The discussion in Section 5 examines some further issues and avenues for future research, followed by Section 6, the conclusion.

2. Background

2.1 Linguistic background

The distinction between consonants and vowels is a basic one for linguistic phonetics and phonology, despite the existence of borderline cases: consonants which can serve as syllabic centers (liquids like *r* and *l*) and glides (*w* and *y*) that switch easily between these categories in morphological alternations.

In Harris's *Structural Linguistics* (Harris 1951[1946]), the consonant/vowel distinction makes its first appearance in section 7.21 (page 61), a section that deals with stating the environments of linguistic segments. A footnote says "V indicates any segment of a group which we call vowel. In most languages it is convenient to set up this group, on distributional grounds, in contrast to consonants (C)".

It is our purpose to specify those grounds, though the exact nature of the contrast must first be clarified. Both the definition of the contrast and the phonological level to which it is applicable must be specified.

The difference between consonants and vowels has a fairly clear articulatory basis, with consonants having a stopped or restricted air flow through the vocal cavity, while vowels do not, though specifying the exact acoustical or articulatory properties of the contrast is difficult. Contrasts in phonology are most often encoded by means of distinctive features. A distinctive feature is a distinction (defined by means of acoustical or articulatory properties, or both) that divides linguistic sounds into two categories, characterized by the presence or absence of the feature. A plus sign [+] indicates the presence of the feature, while a minus sign [−] indicates its absence. For a more detailed discussion see Hyman (1975) chapter 2.

Jakobson (Jakobson & Halle 1956) proposed that vowels and consonants are distinguished by means of two features: *vocalic* and *consonantal*. Consonants are *+consonantal* and *–vocalic*, while vowels are *–consonantal* and *+vocalic*. (Glides, such as *w* and *y*, are *–consonantal* and *–vocalic*, while liquids, such as *l* and *r*, are *+consonantal* and *+vocalic*.) Using this set of features, the contrast between vowels and consonants is between linguistic segments that are [*–consonantal*, *+vocalic*] (vowels) and the remaining three feature complexes (consonants).

The articulatory characteristics that define these features are as follows (Chomsky & Halle 1958): Vocalic sounds are produced with an oral cavity in which the most radical constriction does not exceed that found in the high vowels [*i*] and [*u*]. Consonantal sounds are produced with a radical obstruction in the midsagittal region of the vocal tract. While these definitions make sense for most consonants and vowels, their application to liquids and glides is a bit more questionable. Glides do not have a radical constriction (since they are *–consonantal*), but nevertheless have one that is narrower than that of high vowels (since they are *–vocalic*). It seems questionable whether such a fine distinction could be a natural one. Similarly, liquids would have a radical constriction (since they are *+consonantal*), but one that somehow is less than that of high vowels (since they are *+vocalic*). This combination of constraints seems physically impossible.

Within a 'prime feature' framework, such as that outlined in Ladefoged (1975), it is the feature *Syllabic* that differentiates consonants from vowels. Ladefoged notes, "In the discussion of syllables it was pointed out that there is no agreed physical measurement corresponding to syllabicity. But there is no

doubt that segments can be described phonetically as being syllabic (100 percent) or nonsyllabic (0 percent)."

For these and other reasons, Chomsky & Halle suggest replacing the feature *vocalic* with the feature *syllabic*, which characterizes all segments constituting a syllabic peak. Thus, standard consonants, liquids, glides, and nasals are *–syllabic*, while vowels, syllabic liquids and syllabic nasals are *+syllabic*. This feature fits our purposes. We provide a proposed method for distinguishing *+syllabic* segments from *–syllabic* ones below.

The Chomsky & Halle feature system is also designed to be capable of representing both underlying phonemic contrasts as well as surface distinctions (the result of applying phonological rules). This is useful, since we seek to represent the distinction between *+syllabic* and *–syllabic* segments at a surface level. We choose the surface level because the distinction between vowels and consonants is a basic one, and, as such, it ought to be determinable on the basis of a speech stream that has not been analyzed, whether for the purposes of linguistic analysis or for language learning purposes.

In any case, there does not appear to be a clear and simple way of specifying the consonant/vowel distinction either in terms of articulation or acoustics. It is our goal to provide a distributional test for this distinction.

2.2 Earlier Work

Some 40 years ago, Harary and Paper published an analysis of phoneme frequency in a Japanese text. This 1957 article, "Toward a General Calculus of Phonemic Distribution," described phoneme co-occurrence in abstract relational and graph theoretical terms. Appearing in *Language* just two years after Harris's influential "From Phoneme to Morpheme" (Harris 1955), it suggested a numerical method for describing the distribution of the phonemes in a language. Let f represent a generic phoneme. Working from a corpus of phonemic text, the predecessor set, $P(f)$, and the successor set, $S(f)$, of a phoneme were defined as the sets of phonemes which preceded or followed f in the analysis corpus.

For example, in a short Spanish text such as "*el titulo de la conferencia*" (using letters instead of phonemes and disregarding word boundaries), the predecessor set, $P(e)$, consists of the set $\{d,f,r\}$, while the successor set, $S(e)$, is $\{l,r,n\}$. Similarly $P(t) = \{l,i\}$ and $S(t) = \{i,u\}$.

A *relation*, R, *on a set*, M, is a collection of ordered pairs (x,y) of elements of M. We refer to R as the succession relation on M. In the terminology of graph

theory, x is *adjacent to y* and y is *adjacent from x*. Thus in the preceding paragraph, e is adjacent to (each element in) $S(e)$ and adjacent from (each element in) $P(e)$. We now define some basic properties of relations. We say that

- R is *reflexive* if for each x in Φ, we have (x,x) in R, so that x is in both $P(x)$ and $S(x)$.
- R is *irreflexive* if for each x in Φ, (x,x) is *not* in R, so that x is in neither $P(x)$ nor $S(x)$.
- R is *symmetric* if (x,y) in R entails (y,x) in R, or equivalently, if x in $P(y)$ implies y in $P(x)$.
- R is *transitive* if, for any three x, y, z, distinct, (x,y) and (y,z) in R implies that (x,z) is in R.
- Finally, R is *complete* if for any two elements of F, they are adjacent in the corpus and thus members of R.

In Harary & Paper (1957), M was the set of all phonemes in a given corpus and R contains (x,y) whenever x is followed by y in the corpus. Harary & Paper examined two dialects of Japanese with respect to these properties of their respective relations as defined above. We build on this work by examining the relation R in terms of graph theory.

3. Mathematical background

Formally, a *digraph* (*directed graph*), D, consists of a set V of *nodes* (or *vertices*) and a set A of *arcs* each of which is an ordered pair, (u,v), of two nodes denoting a directed line segment from u to v. In terms of relations, a digraph is thus simply an irreflexive relation. A *graph*, G, consists of a set V of *vertices* or *nodes* and a set E of *edges*, each joining two distinct nodes. Thus an arc is a directed edge. In other words, a graph is a symmetric, irreflexive relation, which is neither empty nor infinite (Harary 1969). While Harary & Paper examined the phonemic digraph of a language corpus, we consider its phonemic graph, in which the edge uv is in E, if either of the ordered pairs (u,v) or (v,u) are in the digraph.

A *partition* of a set, M, is a collection of disjoint (mutually exclusive) subsets of M, none of which is empty, whose union is all of M. We are only considering partitions into two subsets. Then a graph G is called *bipartite* if its node set V can be partitioned into two subsets (U and W) so that no edge joins two nodes in the same subset.

The concept of a bipartite graph is closely related to that of graph coloring. We can regard the nodes in U as having umber color and W having white color. In a bipartite graph, every edge joins two nodes with different colors. The chromatic number of a graph, written $\chi(G)$, is the smallest number of colors that can be assigned to the nodes so that no edge joins two nodes with the same color. Clearly, in a bipartite graph with one or more edges, the chromatic number is 2.

Given an edge $e = uv$, node u and edge e are *incident*, as are v and e. A *walk* of a graph G is an alternating sequence of nodes and edges in which each edge is incident with the two nodes immediately preceding and following it. A *path* is a walk with all nodes distinct, and a *cycle* is obtained from a path with at least 4 nodes, by identifying the first and last nodes. A characterization theorem, due to König (1936), is that a graph is bipartite if and only if every cycle has an even number of nodes.

The symbol $\lceil x \rceil$ is called the *ceiling* of the real number x, and represents the next higher integer to x, as illustrated for the well-known transcendental number $\pi = 3.14159\ldots$, so that $\lceil \pi \rceil = 4$. We note that Harary et al. (1977) defined the *biparticity* of a graph G, written $b(G)$, to be the minimum number of bipartite subgraphs required to cover $E(G)$. They proved there that for any graph G,

$$b(G) = \lceil \log_{(2)} \chi(G) \rceil$$

4. Methods

4.1 Initial work

A decade ago, we wrote a preliminary report (1991) on the bipartite distribution of Hawaiian phonemes. We report briefly on that introductory work here. Language texts can be seen as phonemically bipartite if the phoneme inventory which makes up the text can be divided into two groups so that any pair of sequential phonemes in the text consists of one phoneme from one group and one from the other. Given the natural class distinction of consonants and vowels, a text could be more or less bipartite depending on the percentage of phoneme pairs which violate the bipartite constraint. The bipartiteness of a language correlates with the proportion of CV or VC syllable structures in the language.

Our hypothesis, borne out by a short empirical study of a corpus of Hawaiian proverbs, is that with respect to the consonant/vowel partition,

Hawaiian is quite bipartite, since both consonant clusters and vowel clusters are rare in Hawaiian. As defined above, a graph $G = (V,E)$ with node set V and edge set E is *bipartite* if there is a partition, $V = U \cup W$ such that every edge e of E joins an umber node in U with a white node in W. Now consider a graph $G = (V,E)$ which is not bipartite but has a partition, $V = U \cup W$, in which most, but not all, of the edges join U-nodes with W-nodes. Then it is natural to define the *bipartiteness ratio* of a specific parition of the nodes of G, as the number of uw edges (for u in U, w in W) divided by the total number of edges. In a bipartite graph this ratio is 1.

Given a graph derived from a language corpus, it is feasible to count the number of times a given edge is traversed in the corpus. This count can be called the *weight*, $w(e)$, of the edge e. Given such a weighted graph, the *weighted bipartiteness ratio* of the graph can be defined as the weight of the U–W edges divided by the total weight of the edges. In a bipartite graph this ratio is again 1.

We calculated the bipartiteness ratios of the small corpus of Hawaiian proverbs in Appendix A, treating long vowels and diphthongs as single phonemic elements. (While proverbs may not be most representative in terms of lexical elements or syntactic structure, they are usually representative in terms of phonology.) We looked at both weighted and unweighted graphs. We also looked at two different graphs, one taking each word separately, and one looking at continuous text and ignoring word boundaries. The adjacency matrix for one of the graphs is in Appendix B. The bipartiteness ratio of each of these four graphs, using a consonant/vowel partition is shown in Table 1.

Table 1. Bipartiteness ratios for Hawaiian

	Unweighted	Weighted
Including word boundaries	.85	.95
Ignoring word boundaries	.73	.86

The best result (the highest bipartiteness ratio) was obtained from the weighted bipartiteness ratio, using text that marked word boundaries. This result makes sense in that syllable-structure constraints are generally not enforced across word boundaries, so that one would expect a lower bipartiteness ratio if word boundaries are ignored. Similarly, one would expect consonant and vowel clusters to be more marked and less common than a simple CV syllable structure, so that a weighted calculation should have a higher bipartiteness ratio than an unweighted one.

4.2 Generalization of the hypothesis

In a bipartite connected graph, the division into two sets is unique; in a disconnected bipartite graph this is not so. However, in a connected graph that is not bipartite, any division into two sets has a bipartiteness ratio as defined above. This suggests the hypothesis that the partition of the phoneme set into consonants and vowels produces the most bipartite graph. If this hypothesis is true, then one would have a method for distinguishing between consonants and vowels in a phoneme set by finding the partition that resulted in the highest bipartiteness ratio for the graph.

Intuitively, this hypothesis makes sense. Although most languages are less bipartite than Hawaiian, vowels still form the syllabic center of each syllable. Simple CV structure is, if not the most common, then one of the most common forms of syllable structure. Consonants which occur within or at the beginning of consonant clusters are also likely to occur at least as frequently as the sole onset of a syllable or at the end of a consonant cluster. It is plausible then that on the whole, they would be more likely to be adjacent to a vowel than to occur between other consonants.

To test this hypothesis, we initially wrote a small C program, which computed the four results found in Table 1, given a language corpus C of a language L with n phonemes $(p_1, \ldots, p_n)$, written phonemically, for any two-fold partition of the phonemes of L. Four n by n arrays were constructed. In the first array, the element [i,j] contained the number representing the total number of times the phoneme sequence $p_i p_j$ appeared in C. In the second array, element [i,j] contained 1 when the phoneme sequence $p_i p_j$ occurred in C, and 0 otherwise. In the third array, the array element [i,j] contained the number representing the total number of times the phoneme sequence $p_i p_j$ or (where # represents a word boundary) the phoneme sequence $p_i \# p_j$ appeared in C. In the fourth array, element [i,j] contained 1 if either $p_i p_j$ or $p_i \# p_j$ appeared in C, and 0 otherwise. Arrays two and four thus represent adjacency matrices of digraphs of C. Arrays one and three represent adjacency matrices for *networks* (weighted digraphs) of C. The program then computed four bipartiteness ratios for every two-fold partition of the phoneme set, for each of the four graphs of C based on the four adjacency matrices. The partition having the highest bipartiteness ratio for each graph was stored and printed out at the end of the program, along with its bipartiteness ratio.

We looked at four languages using these programs, starting with Hawaiian. This language, with its limited phoneme inventory and standard CV

syllable structure, provided a good starting point. A large amount of phonemic data is somewhat difficult to obtain, so we also looked at languages which are written rather phonemically (unlike English or French, for example). Our candidates were Esperanto and Spanish. We were also able to obtain some phonemic data from Bruce Nevin for Achumawi, a Native American language of northern California. In preparing a corpus for use with the program, we lower-cased all characters and removed all punctuation, formatting, and numerics from the text. We then phonemicized the text as much as possible, a task which we describe in each section of the results. In most cases, this involved merely substituting an arbitrary single symbol for a sequence of letters representing one phoneme in the language.

5. Results

For each of the four calculations, we show the most bipartite division, and its bipartiteness ratio. We show this ratio in two forms. First we show the actual calculation, that is, the number of bipartite edges (either weighted or unweighted) divided by the total number of edges (again, either weighted or unweighted).

5.1 Hawaiian

For Hawaiian, as indicated above, long vowels and diphthongs were treated as independent phonemes. The glottal stop phoneme is represented by the apostrophe. The corpus size is 285 words (1,118 characters).

Table 2. Hawaiian

Case	W/U	SW/RT	1st Division	2nd Division	Bipartiteness Ratio
1	W	SW	a ā e ē i o ō u ū ae ai ao au	' h k l m n p w	518/546 = .949
2	U	SW	a ā e ē i o ō u ū ae ai ao au	' h k l m n p w	63/75 = .84
3	W	RT	a ā e ē i o ō u ū ae ai au	ao ' h k l m n p w	715/830 = .861
4	U	RT	a ā e ē o ō u ū ae ai au	i ao ' h k l m n p w	73/92 = .793

Note: W/U = weighted/unweighted; SW/RT = single word/running text

Including word boundaries in the count (that is, excluding from the count sequences consisting of a word-final phoneme followed by a word-initial phoneme) improves the division of phonemes along vowel-consonant lines. These results correctly divide phonemes into consonant and vowel groups. (In

the short corpus we used, long *i* [*ī*] does not occur. We therefore do not include it in the division, since as an unconnected node, it could appear in either partition without changing the bipartiteness ratio.)

Disregarding word boundaries produced less satisfying results, since *i* and *ao* were classed with the consonants. Since basic Hawaiian syllable structure is CV, in all instances *ao* appeared following a consonant, but by accident also appeared at the end of a word and before another word beginning with a vowel. This suggests that using a larger corpus might produce better results, even disregarding word boundaries.

5.2 Esperanto

It is not always easy to find phonemic transcriptions of languages, though some alphabets approximate such a transcription more accurately than others (English spelling, for instance, being notably non-phonemic). So, for a second language we selected Esperanto, since its alphabet was designed to be phonemic. Each letter represents one phoneme, though there are several diphthongs, which again we counted as single phonemes.

Esperanto is not nearly so bipartite as Hawaiian, so we thought it would be interesting to see what its bipartiteness ratios turned out to be, and whether our generalized hypothesis held for Esperanto as well. Esperanto in addition has more phonemes than Hawaiian. Esperanto has 35 phonemes, while Hawaiian has 22. This results in a much larger number of possible partitions. The corpus size was 710 words (4,229 characters).

Table 3. Esperanto

Case	W/U	SW/RT	1st Division	2nd Division	Bipartiteness Ratio
1	W	SW	a e i o u au eu aj ej oj	uj b c c d f g g h h j j k l m n p r s s t v z ž	2297/2809 = .818
2	U	SW	a e i o u au eu aj ej oj	uj b c c d f g g h h j j k l m n p r s s t v z ž	129/200 = .645
3	W	RT	a e i o u au eu aj ej oj	uj b c c d f g g h h j j k l m n p r s s t v z ž	2707/3518 = .769
4	U	RT	a e i o u au eu aj ej oj uj	b c c d f g g h h j j k l m n p r s s t v z ž	150/247 = .607

For Esperanto, only one of the calculations produced an exact division between vowels and consonants, the one which looked at the unweighted graph that ignored word boundaries. Based on intuitive predictions, this graph would be the least likely to produce the correct division. The phoneme in question is the diphthong *uj*, which in our corpus appears only five times. All five occurrences were in correlative pronouns, in which the diphthong is created by adding the plural marker *-j* to the pronoun. In each case the pronoun ended in *iu*, so that the resulting diphthong *uj* was, in each case, preceded by an *i*. In three of those cases, it was followed by the accusative marker *-n*, while in the remaining three cases it ended the word.

5.3 Achumawi

Thanks to the kindness of Dr. Bruce Nevin, we have a corpus of Achumawi words. These do not form a text corpus, but are simple word lists. Therefore, we did not compute figures for the two graphs which disregard word boundaries. Achmawi has a large phoneme inventory. Vowel length is distinctive, as are high and low tones, resulting in 20 phonemic vowels. Consonant length is also distinctive, and, in addition, there is a distinction between glottalized and unglottalized consonants, as well as between plain and aspirated consonants. This proliferation of distinctive features results in 73 phonemes in the word list.

The large phoneme inventory in Achumawi tested the limits of our program. The program checks every possible partition, and each additional phoneme doubles the number of partitions to check. With 73 phonemes, an exhaustive search was not possible. With a phoneme set this large, other search procedures would be better, such as simulated annealing or genetic algorithms. In this case, we simply ordered the phoneme set with vowels first and then consonants, so that the implemented search procedure would find the consonant/vowel partition fairly early on in the search. While it was not possible to run the program long enough to finish the complete search (it would have required several days to run), we are fairly confident that the results are valid. In the list below, glottalization is indicated by the apostrophe, while length is represented by a following colon. (The /'h/ is an epiglottideal spirant.) Aspirated consonants are represented by the set of voiceless consonants and plain consonants by the voiced equivalent. Vowels with high tone are indicated by capitalization. The corpus size was 1,307 words (12,106 characters).

Table 4. Achumawi

Case	W/U	SW/RT	1st Division	2nd Division	Ratio
1	W	SW	a e i o u a: e: i: o: u: A E I O U A: E: I: O: U:	b b: p p: 'p d d: t t: 't 't: j j: c c: 'c 'c: g g: k k: 'k q q: 'q 'q: x x: m m: 'm 'm: n n: 'n l l: 'l s s: w w: 'w 'w: y y: 'y 'y: h h: 'h 'h:	8477/9492 = .893
2	U	SW	a e i o u a: e: i: o: u: A E I O U A: E: I: O: U:	b b: p p: 'p d d: t t: 't 't: j j: c c: 'c 'c: g g: k k: 'k q q: 'q 'q: x x: m m: 'm 'm: n n: 'n l l: 'l s s: w w: 'w 'w: y y: 'y 'y: h h: 'h 'h:	613/765 = .801

These calculations, though incomplete, produce the appropriate consonant/vowel partition.

5.4 Spanish

Large amounts of online Spanish text are readily available. The difficulty with Spanish, however, is that it is not entirely phonemic. In addition to the standard text preparations, we made the following changes. Accented vowels were replaced by their unaccented equivalents, as the accent indicates stress only and is not distinctive. The pairs of letters *ch*, *rr*, and *ll* were treated as separate phonemes. The letter combination *qu* was treated as the phoneme *k*. The letter *c* was divided between the phoneme *k* (where it appeared before *a*, *o*, *u*, *l*, or *r*) and the phoneme *s* (elsewhere). The letter *h* was eliminated, as it is not pronounced. The letters *b* and *v* were regarded as the same phoneme. We followed the Latin American dialect and regarded the letter *z* as equivalent to the phoneme [*s*]. The two-letter combinations, *ai*, *ei*, *oi*, *ia*, *ie*, *io*, *iu*, *ua*, *ue*, *ui*, *uo*, *au*, *eu*, *ou* were replaced by *ay*, *ey*, *oy*, *ya*, *ye*, *yo*, *yu*, *wa*, *we*, *wi*, *wo*, *aw*, *ew*, *ow* respectively. The resulting corpus is not entirely error free. Abbreviations and foreign words were not removed, for example. In addition, a few of the two-letter vowel combinations should have remained as two vowels. The corpus size was 7,521 words (46,146 characters).

Table 5. Spanish

Case	W/U	SW/RT	1st Division	2nd Division	Bipartiteness Ratio
1	W	SW	a e i o u	y w l r ch rr ll p t k b d g f s m n ñ j x	25889/31099 = .832
2	U	SW	a e i o u y w n l r	ch rr ll p t k b d g f s m ñ j x	130/201 = .642
3	W	RT	a e i o u	y w l r ch rr ll p t k b d g f s m n ñ j x	29711/38619 = .769
4	U	RT	a e i o u y w s x n l	r ch rr ll p t k b d g f m ñ j	137/231 = .593

The Spanish results clearly show the value of a weighted measure. Neither of the unweighted measures produces the correct division, including the glides *w* and *y*, as well as *n* and *l* with the vowels. The unweighted measures also differed in their results, so that the calculation counting word boundaries included *r* with the vowels, while the calculation not counting word boundaries included *s* and *x* instead. In part this may be due to the alternative treatment of diphthongs in the surface representation of Spanish. Instead of treating them as separate phonemic vowels, as was done with Esperanto and Hawaiian, diphthongs were coded as a combination of a vowel with a glide (*y* or *w*) preceding or following. This had the effect of creating more consonant-glide clusters in the corpus than would have been the case if diphthongs had been treated as separate phonemes.

6. Discussion

This technique is a possible tool for linguistic analysis. It might serve also as part of a language-learning process. We would expect this technique to be most useful at an early stage of analysis or language learning. However, two aspects of the technique might appear to contradict that intent.

First we used phonemic data, rather than phonetic data. Surely, a phonemic analysis would require a prior division of language sounds into vowels and consonants. Harris thought so, since he made use of the distinction in determining the phonemic inventory of a language. However, we feel that the technique would work as well using a phonetically-transcribed corpus as a phonemically-transcribed one. In fact, it ought to work better, since allophonic

differences have not been removed from the transcription. If anything, it should be clearer that a single phonetic segment is either vocalic or consonantal. There is a difficulty, however, if phonemes consisting of double segments have not been identified. This would include diphthongs, long vowels (in some cases), and affricates. The Spanish example, however, shows that the technique still works even if diphthongs are encoded as *glide+vowel* or *vowel+glide*, rather than as separate phonemes. In our examples, we used a phonemic transcription because phonetic transcriptions were not readily available in enough quantity to be useful.

Second, we have shown that the technique works best when a corpus is used that contains word boundaries. That raises the question of whether word boundaries can be identified without first dividing the sounds of a language into vowels and consonants. If not, then again the technique is circular — division of sounds into vowels and consonants presupposes the ability to mark word boundaries, which presupposes the ability to distinguish vowels from consonants. We would claim, though, that there is nothing in either Harris's morpheme-boundary discovery algorithm or other morpheme discovery procedures (e.g., Goldsmith 2000) that requires a phonemic corpus. These procedures will operate acceptably on a phonetic corpus, so that, at the very least, morpheme-boundary marking and consonant/vowel partitioning can occur in tandem.

The limited experiments we performed suggest additional areas of research. First, of course, is the task of providing additional verification for the method. Using this technique on other languages, particularly those known to have less of a preponderance of CV syllable structure, is vital. So also is testing the technique on phonetic, rather than phonemic, data. Testing with various methods of phonemic transcription would also be helpful. Other methods of calculation might be investigated, such as calculations based on the directed graph of the corpus, rather than the undirected graph. For example, it might be interesting to determine the bipartiteness ratio using only the successor or predecessor sets of a phoneme. Results from both calculations could be intersected to produce a core set of consonants and vowels.

It would also be of interest to compare the bipartiteness ratios of the consonant/vowel partition with that of other partitions of the same set of phonemes, such as high/low or front/back, or even to examine the distribution of the bipartiteness ratios of all partitions, to see whether the consonant/vowel partition stands out as significantly more bipartite than other partitions.

Finally, there may be other linguistic categories, relationships, or distinctions that could be amenable to discovery by similar techniques, using other

calculations based on different graph properties. Or there may be other applications of this technique outside of linguistic analysis or language learning. Perhaps it could be usefully applied to undeciphered scripts. Or it may be of typological interest to classify languages on the basis of their bipartiteness or on the basis of other properties of their graphs.

7. Conclusion

We have shown that, for a number of languages, the partitioning of the phonemes of those languages into consonants and vowels produces a higher bipartiteness ratio than any other partition. Of four possible ways to calculate the bipartiteness ratio of a corpus, that which uses word boundary markers and weighted edges produces the best results.

Appendix A: Hawaiian text

1. *Nā ʾōlelo no ʾeau*
   "Traditional sayings.''
2. *ʾA ʾole hiki i ka iʾa liʾiliʾi ke ale i ka iʾa nui.*
   "A small fish cannot swallow a big one."
   [A commoner cannot do anything to a chief.]
3. *ʾA ʾole make ka waʾa i ka ʾale o waho, aia no i ka ʾale o loko.*
   "A canoe is not swamped by the billows of the ocean, but by the billows near the land."
   [Trouble often comes from one's own people rather than from outsiders.]
4. *ʾA ʾole no i ʾike ke kanaka i na nani o kona wahi i hānau ʾia ai.*
   "A person doesn't see all the beauties of his birthplace."
   [One doesn't see how beautiful his birthplace is until he goes away from home.]
5. *E hine auane ʾi na nuku, he pōmaika ʾi ko laila.*
   "Where the mouths are shiny (with fat food), prosperity is there."
   The prosperous have the richest food to eat.
6. *E hiolo ana na kapu kahiko*; *e hina ana na heiau me na lele*; *e hui ana na moku*; *he iho mai ana ka lani a e piʾi ana ka honua.*
   "The ancient kapu will be abolished; the heiau and altars will fall; the islands will be united; the heavens will descend and the earth ascend."
   [Chiefs will come down to humble positions and commoners rise to positions of honor.]
7. *E ho ʾi e pe i ke ōpū weuweu me he moho la. E ao o ha ʾi ka pua o ka mauʾu ia ʾoe.*
   "Go back and hide among the clumps of grass like the wingless rail. Be careful not to break even a blade of grass."

[Return to the country to live a humble life and leave no trace to be noticed and followed. Used as advice to a young person not to be aggressive or show off.]

8. *He ali'i ke aloha, he kilohana e pa 'a ai.*
   "Love is like a chief; the best prize to hold fast to."
9. *He 'ālo 'ilo 'i, ka i'a waha iki o ke kai.*
   "An 'Alo 'il 'i, a fish of the sea that has a small mouth."
   [Said of one who always has little to say.]
10. *I 'ike 'ia no ke ali'i, i ka nui o na maka 'āinana.*
    "A chief is known by his many followers."
11. *I kāhi 'e ka mālia, hana i ka makau. I kahiki ka ua, ako 'ē ka hale.*
    "While fair weather is still far away, make your fishhooks. While the rain is still far away, thatch the house."
    [Be prepared.]
12. *I ka holo no i ka alahao a pi'i i ka lani.*
    "While going along the railroad one suddenly goes up to the sky."
    [A drinker soon finds himself 'up in the clouds.' An expression used by sweet-potato beer drinkers.]
13. *I kanaka no 'oe ke mālama i ka kanaka.*
    "You will be well served when you care for the person who serves you."
14. *I kani ko 'aka i ka le 'ale 'a; i pu'u ko nuku i ka huhū; i le 'a ka nohona i ka mā 'ona.*
    "One laughs when joyous; sulks when angry; (is) at peace with all when the stomach is satisfied with food."
15. *I kani no ka 'alae i ka wai.*
    "A mudhen cries because it has water."
    [A prosperous person has the voice of authority.]
16. *I kani no ka pahu i ka 'olohaka o loko.*
    ["It is the space inside that gives the drum its sound."
    [It is the empty-headed one who does the most talking.]
17. *I ka noho pu ana a 'ike i ke aloha.*
    "It is only when one has lived with another that one knows the meaning of love."
18. *No nehinei 'e nei no'; heaha ka 'ike?*
    "(He) just arrived yesterday; what does he know?"
19. *Nui kalakalai, manumanu ka loa 'a.*
    "Too much whittling leaves only a little wood."
20. *Nui pūmai'a 'olohaka a loko.*
    "Large banana stalk, all pith inside."
    [Said of a person with a large physique but with no strength to match it.]
21. *Pa 'a no ka 'aihue i ka 'ole.*
    "A thief persists in denying his guilt."
22. *Pae mai la ka wa 'a i ka 'āina.*
    "The canoe has come ashore."
    [Hunger is satisfied; or, one has arrived hither.]

23. *Pali ke kua, mahina ke alo*
"Back (as straight) as a cliff, face as bright as the moon."
[Said of a good-looking person.]
24. *Pā mai, pā mai ka makani o Hilo; waiho aku i ka ipu iki, hō mai i ka ipu nui.*
"Blow, blow, O winds of Hilo, put away the small containers and give us the large one."
[*La 'amaomao*, the god of wind, was said to have a wind container called *Ipu-a-La 'amaomao*. When one desires more wind to make the surf roll high, or a kite sail aloft, he makes this appeal.]
25. *Pā no, lilo!*
"Touched, gone!"
[Said of one with deft fingers: A touch and the thing is gone!]

Appendix B: Weighted adjacency matrix including word boundaries

The following matrix displays the weighted adjacency counts for the small corpus of Hawaiian proverbs in Appendix A, including word boundaries in the count (that is, excluding from the count sequences consisting of a word-final phoneme followed by a word-initial phoneme).

	a	ā	e	ē	i	ī	o	ō	u	ū	ae	ai	ao	au	ʻ	h	k	l	m	n	p	w	#
a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	9	13	15	1	22	1	0	97
ā	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	3	0	1	0	0	3
e	1	0	0	0	3	0	0	0	2	0	0	0	0	0	5	1	0	1	0	0	0	0	47
ē	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
i	4	0	0	0	0	0	1	0	0	0	0	0	0	1	9	1	9	5	0	4	2	0	66
ī	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
o	1	0	2	0	0	0	0	0	0	0	0	0	0	0	4	8	4	8	0	4	0	0	36
ō	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1
u	4	0	1	0	6	0	0	0	0	0	0	0	0	0	1	1	2	0	1	0	0	1	13
ū	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2
ae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
ai	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1	1	0	2	0	0	10
ao	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
au	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3
ʻ	19	1	2	1	18	0	9	0	2	0	0	3	0	0	0	0	0	0	0	0	0	0	1
h	10	1	8	0	11	0	9	1	4	1	0	0	1	0	0	0	0	0	0	0	0	0	0
k	57	1	15	0	5	0	9	0	6	0	0	1	0	1	0	0	0	0	0	0	0	0	0
l	8	0	13	0	7	0	16	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0
m	8	3	2	0	0	0	2	0	0	0	0	7	0	1	0	0	0	0	0	0	0	0	0
n	27	0	5	0	7	0	13	0	10	0	0	0	0	1	0	0	0	0	0	0	0	0	0
p	4	3	1	0	2	0	0	1	6	2	1	0	0	0	0	0	0	0	0	0	0	0	0
w	5	0	2	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
#	20	0	9	0	43	0	9	1	1	0	0	3	1	1	21	23	65	14	19	30	16	8	0

References

Chomsky, Noam & Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper & Row.

Goldsmith, John. 2000. "Linguistica: an automatic morphological analyzer." In *Proceedings of the 36th Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society.

Harary, Frank. 1969. *Graph Theory*. Reading, MA: Addison-Wesley.

Harary, Frank & Stephen Helmreich. 1991. "On the bipartite distribution of Hawaiian phonemes." MCCS-91–225, Memoranda in Cognitive and Computer Science. Las Cruces, NM: New Mexico State University, Computing Research Laboratory.

Harary, Frank, Derbiau Hsu, & Zevi Miller. 1977. "The Biparticity of a graph." *Journal of Graph Theory*, 1:131–133.

Harary, Frank & Herbert H. Paper. 1957. "Toward a general calculus of phonemic distribution." *Language* 33:143–169.

Harris, Zellig S. 1955. "From phoneme to morpheme." *Language* 31:190–222.

Harris, Zellig S. 1951[1946]. *Methods in Structural Linguistics*. Chicago: University of Chicago Press. (Repr. as "Phoenix Books" P 52 with the title *Structural Linguistics*, 1960; 7th impression, 1966; 1984.)

Hyman, Larry M. 1975. *Phonology: Theory and Analysis*. New York: Holt, Rinehart & Winston.

Jakobson, Roman & Morris Halle. 1956. *Fundamentals of Language*. The Hague: Mouton.

Ladefoged, Peter. 1975. *A Course in Phonetics*. New York: Harcourt Brace Jovanovich, Inc.

Peterson, Gordon E. & Frank Harary. 1961. "Foundations of phonemic theory." In *Proceedings of Symposia in Applied Mathematics, Vol. 12: Structure of Language and Its Mathematical Aspects*, 139–165. Providence: American Mathematical Society.

Part V

Applications

Chapter 14

Operator grammar and the poetic form of Takelma texts*

Daythal L. Kendall
Unisys Corporation

1. Introduction

Although not sufficient in and of itself, syntax can reveal much about the artistic structure of a text which is not otherwise easily discernable, and the great power of Zellig Harris's approach to syntax is that it allows the analysis and description of a language on its own terms without forcing it to fit a preconceived mold. The syntactic analysis of Takelma (Kendall 1977) using a Harris-style operator grammar interested and pleased him as further validation of the theory by applying it to " a language unlike Indo-European languages " (p.c.). Also, as Bruce Nevin reported in his obituary of Zellig Harris (Nevin 1992:63), Harris did have an interest in poetry, and it is thus fitting that the theory should, as will be seen below, prove a powerful tool in the analysis and interpretation of texts as poetry.

---

* At the time of first contact with Whites, the Takelma occupied the middle and upper courses of the westward flowing Rogue River in southwestern Oregon. In the mid 1850s, they were removed from their homeland to the Coast Reservation, some to the Siletz Agency, others to the Grand Ronde Agency. In their new environment, they were interspersed among the peoples of more than 20 other groups, all speaking different languages: the United States Government could not have devised a more effective means of destroying these languages and cultures. Further detail on the Takelma can be found in Kendall (1990) and references contained therein. The author would like to express his appreciation for comments and discussion to Regna Darnell, Dell Hymes, Carolyn Kendall, and Kristine Long. However, responsibility for the content must remain with the author.

Although Takelma narratives share some formal techniques, subject matter, and purpose with texts in other indigenous languages of Oregon, Takelma oral literature does not fit neatly into classifications or general descriptions of the literature of that area.[1] Many Takelma stories are oriented toward addressing social issues and teaching appropriate behavior. They deal with social conventions and the consequences of violating those conventions, with world order and the consequences of the disruption of that order. The text analyzed here, which Sapir (1909: 13–17) entitled "Coyote's Rock Grandson," addresses issues such as obtaining a wife, establishing and maintaining the proper social and economic relationships between a man and his wife's parents, and behaving appropriately toward one's neighbors. In it, Coyote is neither trickster nor transformer. Rather, he is an ordinary resident of an ordinary village who, along with his wife and daughter, becomes the victim of socially inappropriate behavior.

Takelma texts are allegorical in nature, i.e., the storyteller arranges and manipulates symbols in a complex web of interrelationships through his/her control of syntax and other devices in a connected and coherent fashion to entertain and educate. If one compares the style of these allegorical texts with that of the text describing how a Takelma house was built, the former have a more intense style, that is, the ideas and meanings conveyed go far beyond the literal meaning of the text. In that vein, these Takelma texts certainly fit a description of poetry attributed to T. S. Eliot that poetry is ". . . not the assertion that something is true, but the making of that truth more fully real to us." (X. J. Kennedy: 1986: 265). Perhaps we can find in Takelma poetry some words, ideas, and feelings which speak to us morally, philosophically, and artistically. Perhaps, we can regain something of the Takelma world view that died with the last speakers of the language.

1. As Dell Hymes points out (p.c.), the stories told by Oregon peoples often "depicted an original inadequate state [of the world], from which a satisfactory state was brought about." The stories often explain why things are the way they are. For example, the Kalapuyan stories from the Willamette River Valley (see Jacobs 1945), took place in a time when all the animals were people, and Coyote is typically the main character playing the roles of trickster, transformer, and/or object of ridicule. Kalapuyan stories are very much concerned with explaining the *why* and/or *how* of the world in which the Kalapuyan people lived. In contrast, Takelma narratives seem less concerned with such topics (although some such as "*Taltal* the Transformer" are) and much more concerned with social issues.

2. The Takelma text as poetry

In the Takelma world, in the normal course of events, a man obtained a wife by paying a negotiated bride price. The husband also had obligations to his wife's parents in the form of gifts after the marriage and payment upon the birth of the first child (Sapir 1907:275). In "Coyote's Rock Grandson," the bride was stolen. Therefore, the culturally correct social and economic relationships with the bride's parents were not established, and other deviations from social norms ensued.

My first explorations of Takelma texts as poetry were inspired by Dell Hymes' early work in this area (see, for example, Hymes 1979, 1981). Using clues such as overt surface markers, parallel constructions, repetitions, etc., he found in some Takelma texts a definite structure (p.c.). Given these preliminary findings, I began to examine some Takelma narratives using methods of syntactic analysis which I learned under the tutelage of Zellig Harris, and found the texts to be highly structured. However, the recovery of the poetic structure of Takelma texts is no simple matter: one must be very careful that one does not force the language to fit some predefined mold.

Given the syntactic analysis of a text, one must also consider other factors. In "Coyote's Rock Grandson", the number three is clearly a recurrent theme in terms of overt markers as well as the syntactic structure:[2]

- Each of the major divisions is tripartite (see the discussion of *Onset*, *Ongoing*, *Outcome* below).
- Coyote's daughter is mentioned three times explicitly (though never by name), and then the Otter youth stole her.

---

2. There is the problem of *what* and *how* to count: Hymes (1990, 1998, et al.) finds in Takelma texts that four is significant. He states (p.c.), "To be sure, four is the salient pattern number in Takelma, so far as is known." and goes on to ask, regarding explicit repetition in threes, "Why is it not the case that the overall relation is fourfold? 1 2 3 leading to 4?" Hymes is coming to the analysis from a different direction than I, and as Regna Darnell indicated (p.c.), each of us can shed some light on the subject without invalidating what the other has done. Given the recurrence of three and seven in my analysis of this text (and others) using syntax, and given the apparent importance of three and seven in items of material culture (see discussion below), my preference is to think in terms of threes and sevens. The importance of these numbers was thus manifested in at least two ways in the culture, and my gut feeling is that this state of affairs is no accident.

- Rock Boy says "I shall go" three times and begins his journey.
- On his journey, he reaches three houses. Incidents at the first and second are part of the annealing process in which he is prepared for his future. Arrival at the third house is a significant event: Rock Boy has reached the house of his maternal grandparents, the only one of the three incidents introduced by "Night it became," a phrase which often signals the end of one phase and the onset of another.
- After Rock Boy entered the third house, *alxìik* "he saw him/her" is used three times as he looks at Coyote and Crane, and then he expresses his belief that he has found his maternal grandparents.
- Three times it is stated that Coyote killed large deer but was deprived of it/them. Then Rock Boy learns of his grandfather's plight and corrects the situation.

The above list is not exhaustive, but it gives an indication of the pervasive influence of *three* in this narrative.

The number *seven*, like *three*, seems to be a factor in Takelma texts. Using the analytical techniques as discussed below in the section on grammar, "Coyote's Rock Grandson" is found to consist of seven main sections which I have labeled "I" through "VII".[3]

In other areas of Takelma culture, the numbers *three* and *seven* seem to have been very important as well. In the heart of the Takelma homeland, the Jacksonville (Oregon) Historical Society had (when I visited the museum in 1975) a number of baskets which are generally considered to be of Takelma (some possibly of Klamath) origin. In most, the primary decorative motif appears three times, and the decoration is usually in three bands. The band containing the primary motif is the widest. In some baskets, there is also a secondary motif which appears in narrow bands seven or fourteen (i.e., seven pairs) times.

---

3. Of the two texts analyzed in Kendall (1980), "Coyote and Pitch" also is divided into seven main sections each of which exhibits the *Onset, Ongoing, Outcome* pattern discussed below. The other, "Coyote Goes Courting", I originally divided into nine sections, but further examination leads me to believe that it too consists of seven major divisions. This reanalysis is supported by the *Onset, Ongoing, Outcome* patterning which is not apparent in the nine sections but becomes clear when the poem is reanalyzed into seven parts. I find this to be the trend in these three poems, and it will be interesting to discover whether these patterns appear in other Takelma narratives.

Looking at the text in detail, with the exception of section I (a traditional introduction), each major section is further divided into 3 subsections designated "a", "b", and "c" following the Roman numeral. This text is lacking what may be a more traditional closing such as that in (22).

(22) *kweltì paapiʔt' lèep'lap'*
finished your *paap* seeds collect and eat them
"It is finished. Go gather and eat your *paap* seeds!"

Such a closing does not appear in this text perhaps because its inclusion would have added an eighth section. If we look at section II, which is analyzed below (see Grammatical Analysis), we find that the three subsections can be viewed in an *Onset Ongoing Outcome* (*OOO*) framework in much the same way as observed previously,[4] that is, as *Onset* (II.a), *Ongoing* (II.b), and *Outcome* (II.c). In fact, each of the major sections II through VII follows the same pattern internally, and the poem as a totality can be analyzed in the *OOO* framework where II is the *onset*, III through VI the *ongoing*, and VII the *outcome*. Further, the situations described in III and IV (which are contemporaneous, but geographically distinct) are the direct result of activities in II, and as they develop, the stage is set for the action in section V. The *Outcome*, V.c., leads directly into the *Onset*, VI.a, and similarly for VI.c. into VII.a. Section VII, which concludes the text, brings all the pieces together as social order is restored.

There are a number of symbols used in the text. Among the more obvious are *black clouds* which portend evil in many cultures and *stone* which is often associated with great strength or invincibility. A little less obvious is the use of *deer*. In Takelma culture, real men killed large deer; anyone who brought home a fawn was something less and certainly not worthy of being called a man. The use of *canoe* is more complicated. The fact that canoes were used for river travel implies that these were generally considered well-to-do people since poorer people of lower social standing often used log rafts. More importantly, the canoe was the instrument by means of which the girl was stolen, social conventions violated, and the social fabric rent. In the end, the canoe brought the one who restored social order. Finally, kinship, particularly the father-son relationship, is a very important symbol. The father may appear only briefly or not at all. It may be that some evil befell the father and the son seeks vengeance or, as in the current discussion, the father had acted inappro-

4. As noted in both Kendall (1980) and Hymes (1981).

priately and disrupted the social order. In "Coyote's Rock Grandson," it is almost as if there is some external force compelling Rock Boy to action, compelling him to seek out his maternal grandparents and, as a result, leading him to correct the situation created by his father.

3. Grammatical analysis

From a grammatical perspective,[5] this analysis of "Coyote's Rock Grandson" is an outgrowth of Kendall (1977), which is based on theoretical work of Zellig Harris (1968, 1975, 1976a, 1976b, p.c.). Of particular relevance for the syntactic description of Takelma are concepts such as entry order, dominant order, entry address, and inequalities of likelihood of occurrence (cf. Kendall 1977: 1–18). For the current discussion, entry order is of particular importance since the textual analysis consists of unraveling and reversing the ordered entry of operators on their arguments. Entry order may be most clearly demonstrated when one sentence contains another as a proper part, for example, sentence (1) is contained in, but is not identical to, sentence (2).[6]

(1) *wiham kwitì naʔnakát nakàspi*
my.father how do.ASSERT.3OBJ.2SG-SUBJ say.ASSERT.2SG-OBJ3-SUBJ[7]
"My father says to you 'How did you do to them?'" (74.11–12)

(2) *wiham kwitì naʔnakát nakàspi nakàihiʔ*
my.father how do.ASSERT say.ASSERT say.ASSERT
3-OBJ.2-SG-SUBJ 2-SG-OBJ.3-SUBJ 3-SUBJ.QUOT
"My father says to you 'How did you do to them?' she said, it is said."[8] (71.7–8)

---

5. Those who are not interested in the grammatical analysis may skip over this section without handicap.

6. References for individual Takelma examples are taken from Sapir (1909) unless otherwise specified and will be given in the form "page.line(s)", for example 74.11–12 indicates page 74, lines 11 through 12. Details of morphology will not be discussed here since these may be found in Sapir (1922) and, to some extent, in Kendall (1977).

7. ASSERT, that is *assertive*, is used for the verb forms which Sapir called *aorist*. The verbs so labeled denote not only past tense but also present and sometimes the immediate future. In using the assertive (or aorist), the speaker expresses direct knowledge of a reported event.

8. The element *-hiʔ* which I label QUOT (i.e., quotative) in the analysis and which Sapir translated as "it is said" is rather mysterious. Sapir never really figured it out, and I certainly haven't.

Here, *nakàihiˀ* cannot by itself constitute a sentence. In many cases when an operator becomes the argument of a later entry, the earlier entry is marked to indicate its new status and function, that is, the entering operator imposes an argumenthood indicator on the earlier entry as in (4) where (3) appears in nominalized form as an agentive because of its status as first (and in this case only) argument of the verb.

(3) *tˀemeyanàuˀ* (148.6)
woman-goes-to-get-married. ASSERT.causative.3OBJ.3SUBJ
"They went with her to get her married."

(4) *kanehiˀ alxalií tˀemyànwacˀ*
then.QUOT remain.ASSERT.3SUBJ woman-goes-to-get-married
causative.3-OBJ.agentive
"Then, it is said, the goers-with-her-to-get-her-married remained."
(150.18–19)

The discovery of the underlying poetic structure of a text is the result of reversing the process of formation, sometimes through the use of clues such as the agentive argumenthood indicator. Similarly, (5) receives an argumenthood indicator (ignoring the metatextual operator *kanehiˀ*) and functions as the second argument (or object) of the main verb in (6) where the first argument (or subject) has been zeroed.

(5) *kanehiˀ* [*yulum*] *kuúxtakwa wàata woók*
then.QUOT [eagle] wife.3POSS.REFLX her.to arrive.ASSERT.3SUBJ
"Then, it is said, [Eagle] reached his own wife." (138.3)

(6) *ani kelkuluuk kuuxta wata wuuxta yuulum*
not want.ASSERT.3SUBJ wife.3POSS her.to arrive.3poss eagle
"He did not want Eagle's reaching his wife." (St. Clair 1903–1904, F56)

Given sentences such as (7), one could argue that both "coyote" and "return" could be either operator (later entry) or argument (earlier entry) since there is no overt argumenthood indicator attached to either.

(7) *mii yewèiˀ Skìsi*
now return.ASSERT.3-SUBJ coyote
"Now Coyote returned." (56.7)

(8) *kanehiˀ yewèiˀ pˀelxàcˀ*
then QUOT return.ASSERT.3SUBJ go.out.to.war.AGENTIVE
"Then, it is said, the goers-out-to-war returned." (128.13)

However, the agentive form of the verb *p'elèxaʔ* "they went out to war" appears in the same position and has the same function in (8) as *skìsi* "coyote" does in (7). Since the subject of the verb *yewèiʔ* "they returned" is a nominalized sentence, it must of necessity be the earlier entry, that is, the verb *yewèi*-"return" is the later entry on *p'elxàcʔ* "goers-out-to-war". Although not essential to either theory or description, elegance demands that *skìsi* "coyote" be treated similarly, that is, *yewèi* "return" is the later entry (or operator) on *skìsi* "coyote". This principle can be extended to sentences with both the traditional subject and object functions filled by simple nouns. Additional discussion of operator-argument relationships in Takelma does not seem necessary here, particularly since one can consult Kendall (1977) for more on this topic.

An analysis of the entire text probably would be tedious and boring to read, but a few lines are analyzed here in illustration. The following may be found as lines 8 through 20 (section II) of the text below. In the interest of readability and intelligibility, the analysis is not shown in finest detail, and therefore some words shown as operators are actually the resultants of earlier entries and reductions. The metatextual operator *kanìhiʔ* often marks the beginning of a major division, but not all major divisions necessarily begin with *kanìhiʔ*. In the analysis of the more complex sentences, numbered pairs of parentheses are used to improve readability. Each of the individual sentences in section II is analyzed in detail in (9) through (20).

I have not attempted to identify metalanguage vocabulary in Takelma but use appropriate English words, such as *Sameness* and *Zero* instead. The *Sameness* operator asserts the identity of the two occurrences of *altkem* "clouds" and the *Zero* operator deletes the second occurrence, thus establishing the surface modifying relationship between "black" and "clouds" in (9b). *Simultaneity* leaves its trace in the subordinating suffix *-taʔ*. In (10b), the zeroing of *ais-* 'possess' is the source of the possessive suffix on *tukwì*-"skirt". *Sameness* and *Zero* apply in turn to both "skirt" and "girl" resulting in the reflexive marker being suffixed after the possessive as well as the zeroing of repeated elements. At this point, the complexity of the text begins to build as *Sameness* and *Zero* enter on (9b) and (10b) zeroing *waiwìi* "girl" as shown in (10c). The process of text building continues in (11c) where *Sameness* and *Zero* result in the zeroing of *waiwìi* "girl" in (11b).

(9) a. *kanìhiʔ haaií altkem paatinìʔx tahoóxa*
"Then, it is said, black clouds spread out in the evening
waiwìi pʔakàitaʔ.
the girl, when she was bathing."

b. *Simultaneity* > ($_6$*tahoóxa* > ($_5$ *Zero* > *Sameness* > ($_3$($_1$*paatin-x-* >
in the evening spread out
haaií)$_1$, ($_2$ *altkem* >*haaií*)$_2$))$_3$))$_5$, ($_4$ *pʔakài-* > *waiwìi*)$_4$)$_6$
clouds black clouds bathe girl

(10) a. *tukwìtkwa paixotòxat.*
Her own skirt, she took it off.
b. *Zero* > *Sameness* > (*Zero* > *Sameness* >
(*Zero* > *ais-* > (*waiwìi, tukwì-*)), (*paixotòxat* > (*waiwìi, tukwì-*)))
possess girl skirt take.off girl skirt
c. *Zero* > *Sameness* > ((9b), (10b))

(11) a. *pʔakàiʔ.*
She bathed.
b. *pʔakài-* > *waiwìi*
bathe girl
c. *Zero* > *Sameness* > ((10c), (11b))

While each of the sentences has as its subject *waiwìi* "girl", only (9) retains its subject in non-zero form. The three sentences are bound together as subsection II.a:

II.a. *Zero* > *Sameness* > ((*Zero* > *Sameness* > ((9b), (10b))), (11b)).

The sentence in (12a) begins subsection II.b. As is typically the case with a new section or subsection, the protagonist is different from that of the immediately preceding division, and the first sentence is the only one with a non-zero subject. We know that (12a) and (13a) are separate statements with "otter youth" in apposition to "one" because of the position of "one" relative to "otter youth". If we were counting "otter youths", the numeral would follow the noun being counted. The identity of "one" and "otter youth" is asserted in (13c). The entry of *Sameness* and *Zero* in (14c) zeros the subject "one" of (14b), and similarly for each occurrence of "one" in (15c)–(18c).

(12) a. *ei sìlnakàiʔ mìiʔskaʔ*
One came paddling a canoe,
b. *nak-* > *sìl-* > *mìiʔskaʔ, ei*
do paddle one canoe

(13) a. *pùmxi tapʔaalàu.*
[he was] an otter youth.

b. *tapʔaalàu* > *pùmxi*
young otter
c. *Sameness* > (12b), (13b)

(14) a. *ei paisilìxkwa.*
The canoe he brought to shore.
b. *pai-* > *sil-* > *mìiʔskaʔ, ei*
out.of.[water] paddle one canoe
c. *Zero* > *Sameness* > (13c), (14b)

(15) a. *mii hoyooií waiwii.*
now he.stole the.girl
b. *mii* > *hoi-* > *mìiʔskaʔ, waiwii*
now steal one girl
c. *Zero* > *Sameness* > (14c), (15b)

(16) a. *yaánkw.*
He took her with him.
b. *-kw* > *yank-* > *mìiʔskaʔ, waiwii*
with take one girl
c. *Zero* > *Sameness* > (15c), (16b)

(17) a. *miihiʔ tàn paʔilelèkʔ.*
Now, it is said, a stone he took up and put in her.
b. *paʔ-* > *-i-* > *lelèkʔ-* > *mìiʔskaʔ, tàn, waiwii*
up by.hand put one stone girl
c. *Zero* > *Sameness* > (16c), (17b)

(18) a. *hawilìtkwa kiníkw.*
Into his own house he came with her.
b. *Zero* > *Sameness* >
ha- > (*-kw* > *kiník-* > *mìiʔskaʔ, waiwii*), (*ais-* > *mìiʔskaʔ, -wilì-*)
into with take one girl possess one house
c. *Zero* > *Sameness* > ((17c), (18b))

We can show the structure of this subsection in abbreviated form:

II.b. *Zero* > *Sameness* > ($_6$($_5$*Zero* > *Sameness* > ($_4$*Zero* > *Sameness* > ($_3$*Zero* > *Sameness* > ($_2$*Zero* > *Sameness* > ($_1$*Sameness* > (12b), (13b)$_1$), (14b)$_2$), (15b)$_3$), (16b)$_4$), (17b)$_5$), (18b)$_6$)

Taking this one step further, *Sameness* enters to assert the identity of *waiwii* "girl" in II.a and II.b giving:

II.ab. *Sameness* > (II.a, II.b)

The final subsection is quite short, but formally sufficient, and as was seen above in the discussion of the text as poetry, it is formally and artistically essential as the Outcome in an *OOO* framework. As with subsections II.a and II.b, *Sameness* asserts the identity of *waiwii* in II.a and II.c, and section II is now seen to be a unit.

(19) a. *waiwii mʔʔhwìiʔ.*
"The girl was pregnant."
b. *mʔʔhwìiʔ* > *waiwii*
pregnant girl

(20) a. *hàapxwii pʔaimacʔák.*
"A child she bore."
b. *pʔaimacʔák* > *waiwii*, *hàapxwii*)
give birth to girl child
c. *Zero* > *Sameness* > ((19b), (20b))

The metatextual operator *kanìhiʔ* enters as an overt marker of the beginning of a major division, the overall structure of which is summarized in (21).

(21) *kanìhiʔ* > ($_3$*Sameness* > ($_2$ *Sameness* > ($_1$II.a, II.b)$_1$)$_2$, II.c) $_3$

4. Coyote's Rock Grandson[9]

		Takelma text	English translation
I.		*paáxtis.*	Wolf.
		huulk'.	Panther.
		wili íxtiil.	Houses ten.
		skísi.	Coyote.
	5	*meéx, skísi kuúxta.*	Crane, Coyote's wife.
		peyán mìiskaʔ,	One daughter.
		tiiheliyaʔ wàiwii, skìsi peyàn.	The girl sleeps on a board platform, Coyote's daughter.

9. In giving the English equivalents of the Takelma, I have attempted to convey something of the flavor of Takelma expression. As a result, some of the English is not that which one would normally hear in speech or see in prose but might encounter in poetry.

II.a.		*kanìhiʔ haaií altkem paatinìʔx tahoóxa*	Then, it is said, black clouds spread out in the evening
		waiwìi pʔakàitaʔ.	the girl, when she was bathing.
	10	*tukwìtkwa paixotòxat.*	Her skirt she took off.
		pʔakàiʔ.	She bathed.
II.b.		*ei sìlnakàiʔ mìiʔskaʔ pùmxi tapʔaalàu.*	In a canoe one arrived, an otter youth.
		ei paisilìxkwa.	With the canoe he landed.
	15	*mii hoyooií waiwii.*	Now he stole the girl.
		yaánkw.	He took her with him.
		miihiʔ tàn paʔilelèkʔ.	Now, it is said, a stone he took and put in her.
		hawilìtkwa kiníkw.	Into his own house he came with her.
II.c.		*waiwii məʔhwìiʔ.*	The girl was pregnant.
	20	*hàapxwii pʔaimacʔák.*	A child she bore.
III.a.		*kanìihiʔ skìsi wàiwii hacʔòlol.*	Then, it is said, Coyote, the girl, he missed her.
		òot.	He looked for her.
		tukìi ya tʔayák haxiyá.	Just her skirt he found near the water.
		miihiʔ alpinìx laalí.	Now, it is said, he became one who mourns.
III.b.	25	*ulum pʔìyin mahài tʔomòomt skìsi.*	Formerly, large deer he used to kill, Coyote.
		mii skìsi pʔìyin wetkin	Now, Coyote, deer, he was deprived of them.
		pʔìyax ya okòikin.	Just fawns he was always given.
III.c.		*tkwan kʔemen skìsi.*	A slave Coyote was made.
		ànìiʔ yokʔwooií kwi kiniyakwànmaʔ skìsi pèyan.	Not he knew where she had been taken, Coyote, his daughter.
IV.a.	30	*mii pʔaiyuwòʔ hapxì.*	Now he was born, the child.
		kʔayàiʔ	He grew.
		mii mahài laalì hapxitʔìit'a.	Now big he became, the male child.
		pʔaimacʔák.	She had given birth to him.
		malàk'ihi "kʔasìitʔ hinauú".	She told him, "Your maternal grandparents, up river."
	35	*kanìhiʔ ei wiík'wa.*	Then, it is said, a canoe, he traveled around with it.
		"hindeé, wìkʔasi waata kinàktee."	"Mother! My maternal grandparents, to them I shall go."
		"taʔmàxau."	"Far away."
		"ke kinàktee."	"There I shall go."
		"yalnatàʔ."	"You will become lost."
	40	*"yanàtee. kwinàt'eti."*	"I shall go. How in appearance?"
		"taamolhìt, iicʔòp'al, hatanxmolìt,"	"Red-eared, sharp-handed, in ear red," she
		nakàhiʔ.	said.

		"kʔasìtʔ waiwìit'a pòktan paáls."	"Your maternal grandparent, female, neck long."
		mahài laáli haapxitʔìt'a.	Big had become the male child.
IV.b.		*miihiʔ talyewèiʔ.*	Now, it is said, he went away.
	45	*ei paasaákw.*	A canoe, he paddled it upstream.
		"gungun hàpta yaánt'eʔ," *nakàihiʔ.*	"Otter, his child I go," he said, it is said.
		wìli katak nakàiʔ t'uL t'uL t'uL.	On top of a house he made t'uL, t'uL, t'uL.
		"nekti yaáx wili katák," *nakàiʔ*	"Who graveyard house on top of?" someone said.
		"ke yaáx wili nakàit'eti."	"There's a graveyard house, did you say?"
	50	*"kwinàt'eti texepenát."*	"How in appearance, you who spoke?"
		"maapʔa kwinát'eti eiítp, *kanàt'siʔ eiít'eʔ."*	"Just as you are in appearance, so am I."
		"ne apailìu."	"Well, look inside!"
		apailiwilìuʔ.	He looked inside.
		aliitpàakin.	He was hit.
	55	*siniitkileeʔskwa,* *yoóm menki yàahi laalí.*	He scratched his nose; just full of blood it became.
		apaikinìkʔ.	He went inside.
		aliitpàkatpak.	He hit them all.
		yapʔa heʔiilem	People, he annihilated them.
	60	*yàpʔa tʔomoóm altíl.*	People, he killed them all.
		cʔolx òosip.	"Indian money give me!"
		cʔolx ookoyìn	Indian money he was given.
		tàktakwa kʔowuú.	Over himself he put it.
		kani xi ikìina,	Then water he took.
	65	*alp'ouúp'auhi.*	He blew on it.
		kani pàʔiyeween altíl.	Then he made them all recover.
		cʔolx okoyìn.	Indian money he was given.
		kani yàʔ.	Then he went.
		"gungun hàpta yaánt'eʔ," *nakàihiʔ.*	"Otter, his child I go," he said, it is said.
	70	*kani "nekti yaáx wili katák,"* *nakàiʔ.*	Then, "Who graveyard house on top of?" someone said.
		"ke yaáx wili nakàit'eti."	"There's a graveyard house, did you say?"
		"kwinàt'eti texepenát."	"How in appearance, you who spoke?"
		"ne apailìu."	"Well, look inside!"
		apailiwilókwʔ.	He looked inside.
	75	*aliitpàakin.*	He was hit.
		sinʔiitkileeʔskwa, *yoóm menkìi ya.*	He scratched his nose, blood, just full.
		apaikinìkʔ.	He went inside.
		aliitpàkatpak.	He hit them all.
	80	*heʔiilemèkʔ.*	He annihilated them.
		cʔolx òosip	"Indian money give me,

		tʔüʻüüxtapaʔ.	because you hit me!"
		cʔolx ookoyìn	Indian money he was given.
		xi paayaánkw.	Water, he took it up.
	85	*xi ikìina,*	Water, he took it.
		paayewèiʔ.	They recovered.
		kani yàʔ.	Then he went.
IV.c.		*xùʔn laalí.*	Night it became.
		ei kanau paisaákw.	With the canoe he landed.
	90	*malàk'i k'apàxa,*	She had told her son,
		"ke kʔasìitʔ pòktan paáls,	"There your maternal grandparents, neck long,
		tàamolhít, iicʔòp"al."	red-eared, sharp-handed."
		apaikinìkʔ.	He went inside.
		alxìik taskàxi,	He saw him, long mouthed,
	95	*hatàanxmolhít.*	in ear red.
		alxìik iicʔòp"al.	He saw him, sharp-handed.
		waiwìit'a kaʔal yewèiʔ.	To the female he turned.
		alxìik pòktan paáls,	He saw her, neck long,
		kweélxta paáls.	legs long.
	100	*kati naák'ik wihìnʔá wikʔàsi.*	"So that's what my mother was saying about my maternal grandparents."
		paánx tʔomoók'wa.	Hunger, it was killing him.
		mii xuma òot.	Now food, he searched for it.
		yana tʔayák.	Acorn mush, he found it.
		kʔeleuú.	He slurped it up.
	105	*alxìik kʔàsa.*	He looked at them, his maternal grandparents.
		"wikʔàsi, wihin melèxinaʔ	"My maternal grandparents, because my mother told me
		"iicʔòp'al" nakàitaʔ,	"sharp-handed" since she said,
		"kʔasa pòktan paáls" nakàitaʔ.	"Maternal grandmother, neck long," since she said.
V.a.		*mìihiʔ tʔayák.*	Now, it is said, he had found them.
	110	*k'wàax.*	She woke up.
		"kii eiít'eʔ, kʔasaá!"	"It is I, Maternal Grandmother!"
		"paáxtis hàpxta mìiʔwa," nakàiʔ.	"Wolf, his child probably," she said.
		"paaʔiiyuwunìʔn, iik'wàakwiʔn."	"I'll arouse him. I'll wake him up."
		skìsi mii k'wàax.	Coyote now awoke.
	115	*"kʔasaá, kii eiít'eʔ!*	"Maternal Grandfather, it is I!
		"paateép kʔasaaá!	"Get up, Maternal Grandfather!
		"paánx tʔüm "üüxi.	"Hunger is killing me.
		"yana loóp!	"Acorns, pound them!
		"alhùiʔ, kʔasaá!	"Go hunting, Maternal Grandfather!
	120	*"siíx yàmxta kelkulukwàʔn."*	"Venison, its fat I desire."
V.b.		*skisi pʔiyin mahài tʔomoóm.*	Coyote large deer killed.
		weétkin.	He was deprived of them.

		pʔìyax ka ya okoiíkin.	Fawns, just those he was always given.
		pʔiyìn mahài weétkin	The large deer he was deprived of.
V.c.	125	*lopòxaʔ, yana lopóp.*	She was pounding, acorns she pounded.
		kʔàawant.	She put them into a sifting basket.
		"paihèmk kasàlhi,	"Take it away, quickly.
		pou wetèsinaʔ.	"Soon it will be taken from me.
		paáxtis kuúxta wetèsink."	"Wolf, his wife will take it from me."
	130	*"kii emeʔ eiít"eʔ,*	"I am here.
		wete wetèspikam."	"Not you will be deprived of it."
		xnik kʔemeií.	Acorn dough she made.
		apaihiwilìuʔ.	She ran into the house.
		tan katàk macʔák.	On top of a rock she put it.
VI.a.	135	*miihiʔ paáxtis kuúxta mii weétki,*	Now, it is said, Wolf's wife now she took it from her.
		yana mii weétki.	The acorns now she took from her.
		kèhi yewèiʔ,	There he returned.
		aliitpakàtpok.	He hit them all.
		"kii emeʔ eiít"eʔ.	"I am here!
	140	*"wikʔàsi iitkwanyèekit."*	"My Maternal Grandmother, you have enslaved her."
		altii tʔomoóm.	He beat them all.
		alti k'ailàap'a tʔomoóm.	All the women he beat.
VI.b.		*tahoóxa yewèiʔ altíl.*	In the evening, everyone returned.
		skìsi yewèiʔ.	Coyote returned.
	145	*pʔìyax yàahi lapák skìsi.*	Just a fawn Coyote carried.
		pʔiyin mahài tʔomomanàʔ,	Although a large deer he had killed,
		weétkin.	He was deprived of it.
		pʔìyax ka ya okoyìn.	Just a fawn he was given.
		"kʔasaá, kwitì pʔiyìn mahàiʔa."	"Grandfather, where deer, the big one?"
	150	*"weésin."*	"I was deprived of it."
		Aaaa! skisi wàata hapxitʔìit'a.	"Aaaa! To Coyote a boy,
		heʔiilèmʔxam.	He did away with us.
		tʔomoóxam.	He beat us."
VI.c.		*miihiʔ tʔilàap'akan nous lemèʔx.*	Now, it is said, their husbands gathered next door.
	155	*tʔomoóm hapxitʔìit'a,*	They beat the boy,
		alìihitpakàtpok.	but he struck them all.
		kata yeweyákw.	He got even with them.
		altii tʔomoóm yápʔa,	All the people he beat,
		hapxitʔìit'a xepèʔn.	the boy did so.
	160	*hapxitʔìit'a tʔomùxaʔ.*	The boy was a beating.
		tan hapxitʔìit'a kasiʔ kaʔál niiwàn.	Rock Boy, because of that he was feared.
		yapʔa mahài tʔomoóm tan hapxitʔìit'a.	The big people, he beat them, Rock Boy.

VII.a.		*heʔne nou yewèiʔ,*	Then down river he returned.
		nixa wàata yewèiʔ.	To his mother he returned.
	165	*"alxìikiʔn wikʔàsi.*	"I have seen my maternal grandparents.
		"paáxtis iitkwanyèek'ok',	"Wolf seems to have enslaved them.
		"xùma àlti wetèk'ikam,	"All their food they seem to have been deprived of.
		"pʔii wetèk'ikam" nakàihiʔ,	"Firewood they seem to have been deprived of," he said, it is said.
		nixa kwenhekwàakwanhi.	To his mother he related it.
VII.b.	170	*skisi peyàn kanii yàʔ maxa wàata.*	Coyote's daughter then went to her father.
		p'im ìtepüʔ tʔìit'wi yàʔ.	With a canoe-full of salmon, her husband went.
		moťwók pòmxi,	Otter went as a son-in-law.
		p'im ìtepüʔ yaánkw.	A canoe-full of salmon he took along.
		pùmxi kuúxtakwatìil p'im ìtepüʔ yaánkw,	Otter with his own wife a canoe-full of salmon took.
	175	*maxa wàata apaiwoók.*	Her father, at his house they arrived.
VII.c.		*skìsi kuúxtakwatìil tiihiliíkw pean yewèitaʔ.*	Coyote and his wife were glad When their daughter returned.
		kanii noóu yewèiʔ.	Then down river they returned.

5. Conclusion

Nowhere in the text are the proper behaviors explicitly named or described, but their existence and validity are communicated by means of the story, which illustrates the consequences of violating social norms. It is not a statement that the specific events of the story will necessarily be the outcome in every such situation. Rather, the events are illustrative of the kinds of social disruption that can occur when conventions are violated. "Coyote's Rock Grandson" makes these culturally specific truths real in a way that would not be possible through simple enumeration of principles and rules.

We have seen that "Coyote's Rock Grandson" is highly structured under the influence of the numbers *three* and *seven*. The structure may not conform to what we expect in poetry, but the tripartite structure of *Onset*, *Ongoing*, *Outcome* contrasts very sharply in style and structure with simple narrative. "How a Takelma House Was Built" structurally has little in common with "Coyote's Rock Grandson." Upon inspection of the two texts, one can easily perceive that there is a difference in style, and syntactic analysis using operator

grammar is the very powerful tool which helps us to understand and describe that difference.

The allegorical nature of the Takelma narratives and the communication of ideas beyond the literal meaning of the texts coupled with the formal structures found in them certainly, it seems to me, place Takelma texts squarely in the genre of poetry.

References

Bergman, David, & Daniel Mark Epstein.1983. *The Heath Guide to Poetry*. Lexington, MA: D. C. Heath & Co.

Harris, Zellig S. 1968. *Mathematical Structures of Language*. Interscience Tracts in Pure and Applied Mathematics, no. 21. New York, London, Sydney, and Toronto: John Wiley.

Harris, Zellig S. 1975. *Notes du cours en syntaxe*, ed. by M. Gross. Paris: Le Seuil.

Harris, Zellig S. 1976a. "A Theory of Language Structure." *American Philosophical Quarterly* 13: 237–255.

Harris, Zellig S. 1976b. "On a Theory of Language." *The Journal of Philosophy* 73: 253–276.

Harris, Zellig S. 1982. *A Grammar of English on Mathematical Principles*. New York: John Wiley & Sons.

Hymes, Dell H. 1979. Review of *Giving Birth to Thunder, Sleeping with his Daughter: Coyote Builds North America* by Barry Holstun Lopez. *The Western Humanities Review* 33: 91–94.

Hymes, Dell H. 1981. "*In Vain I Tried to Tell You": Essays in Native American Ethnopoetics*. Philadelphia: Univ. of Pennsylvania.

Hymes, Dell H. 1990. The Discourse Patterning of a Takelma Text: "Coyote and His Rock Grandson." *The Collected Works of Edward Sapir*, VIII: *Takelma Texts and Grammar*, ed. by Victor Golla. Berlin & New York: Mouton de Gruyter. 583–601.

Hymes, Dell H. 1998. *Reading Takelma Texts*. Bloomington: Trickster Press.

Jacobs, Melville. 1945. *Kalapuya Texts. University of Washington Publications in Anthropology*, 11.

Kendall, Daythal L. 1977. *A Syntactic Analysis of Takelma Texts*. Ph. D. Dissertation, University of Pennsylvania.

Kendall, Daythal L.. 1980. "Coyote and Pitch and Coyote Goes Courting." *Coyote Stories II*, ed. by Martha B. Kendall. *International Journal of American Linguistics, Native American Text Series*, 6: 5–45.

Kendall, Daythal L.. 1990. "Takelma." *Handbook of North American Indians: Northwest Coast*, ed. by Wayne Suttles, gen. ed. William C. Sturtevant, vol. XVII, 589–592. Washington, D. C.: Smithsonian Institution.

Kennedy, X. J. 1986. *An Introduction to Poetry*. Boston: Little, Brown & Co.

Nevin, Bruce E. 1992. "Zellig S. Harris: An appreciation". *California Linguistic Notes* 23.2: 60–64 (Spring-Summer).

Sapir, Edward. [1906]. [Field notebooks of Takelma myths, paradigms, and other grammatical notes.] Ms. Pn1.1, Franz Boas Collection of American Indian Linguistics, American Philosophical Society, Philadelphia, PA.

Sapir, Edward. 1907. "Notes of the Takelma Indians of Southwestern Oregon." *American Anthropologist* N. S. 9: 251–275.

Sapir, Edward. 1909. "Takelma Texts". *University of Pennsylvania Anthropological Publications of the University Museum*, 2.1: 1–263.

Sapir, Edward.1922. *The Takelma Language of Southwestern Oregon. Handbook of American Indian Languages*, ed by Franz Boas. Bureau of American Ethnology Bulletin 40.

St. Clair, Harry H. 1903–1904. [Takelma vocabulary and myths.] Manuscript 1655, Washington, D. C.: National Anthropological Archives, Smithsonian Institution.

Chapter 15

A practical application of string analysis*

Fred Lukoff
University of Washington

1. Introduction

Intermediate students of the Korean language who are reading literary and documentary texts sometimes find it difficult to perceive the syntactic structure of a written sentence, especially a longer one, as they look at it on the page. This is because the basic structure, the subject–(object)–verb configuration, is often obscured by modifying adjunct material, which may itself be rather complex. When I was teaching, I would advise a student who seemed to be baffled by a sentence whose structure was not immediately apparent to him to try to pick out the heads of noun phrases and verb phrases while ignoring for the moment the modifying adjunct material.

The approach I would suggest was the method developed by Z. S. Harris in his monograph *String Analysis of Sentence Structure*.[1] By excising the secondary, modifying material of a sentence, the primary material, that is, the elementary sentence, stands out as its structural framework.[2] Korean sentences seem to be well-suited for string analyis because of their linear structure. As Harris says, "The process of determining what is the elementary sentence and what are adjuncts and to what these adjoin . . . is not hard to grasp".[3] My students did indeed take easily to the idea of string analysis, for it provides a

---

* Some of the material and methodology in this chapter appeared in Lukoff (1985) but is presented here in a different way. Benjamin Lukoff gave me invaluable help in the preparation of this chapter. [Many thanks to Benjamin for his continued assistance after his father's death. — Ed.]

1. Harris (1962).
2. Harris (1962: 9–10).
3. Harris (1962: 26).

straightforward, methodical approach to highlighting the basic structure of sentences. Students would carry out these procedures not on paper but in their head, using their knowledge of basic Korean grammar and syntax. It was usually sufficient for them to carry out the analysis down only to the level of relatively large segments of the sentence, to the point where they could recognize the basic structure of the sentence before them.

The method of string analysis proved to be not only effective, in that its results quickly highlighted the basic structure of a difficult sentence, but also efficient in that it did so at a relatively low cost in effort. My purpose here is to demonstrate this practical application of string analysis to Korean written sentences.

Section 2 demonstrates the procedures of string analysis with a shorter sentence and then with a longer one. Section 3 demonstrates how constituent segments and their relations in longer sentences are determined with the aid of string anlaysis. Section 4 shows how string analysis can be useful for the understanding of short sentences as well as longer ones. In these demonstrations, analysis is carried out to some detail in order to make the process clearer and more understandable to readers not familiar with the Korean language. In a few cases, I show how misconstruing the structure of a sentence leads to a mistranslation of it and how applying string analysis can help the student avoid such errors.

2. Steps in the string analysis of Korean sentences[4]

2.1 Sentences containing modifying adjuncts to nouns

Modifying adjuncts are always to the left of the segment to which they apply. Adjuncts to nouns include other nouns, nouns with certain suffixes or particles, and adnominalized predicates. Modifying adjuncts to verbs include noun phrases with certain suffixes and adverbs.[5] Adnominalized predicates are

4. Examples are from various published sources; see Lukoff (1982) for citation of their sources. Korean sentences are in phonemic transcription. See the table of English phonetic approximations at the end of the chapter.

5. The major word classes occurring in the example sentences are: noun (N), verb (V), adverb (ADV), whether primitive or derived. The major sub-classes of verbs are action verbs and adjective verbs, but these sub-classes are not indicated as the distinction is not relevant in this presentation.

one of the two main kinds of adjuncts that tend to hinder students from directly perceiving the basic structure of a sentence; the other is conjunctive clauses, taken up in 3.1 below.

(1) *nanŭn ilŭl kanŭn uli ttal unnŭn ŏlkulŭl chowahanta.*
I like the smiling face of my daughter, who is losing her baby teeth.

Re-write the words in the sentence as a string of word classes:[6]

(1) a. *na-nŭn i-lŭl ka-nŭn uli ttal un-nŭn*
I.as.for tooth who.is.replacing we daughter which.is.smiling
N-TOP N-OBJ V-ADN N N (NP$_1$) V-ADN
ŏlkul-ŭl chowaha-nta
face like
N-OBJ (NP$_2$) V-S.C.

(i) A scan of (1) shows that it contains two NPs, underlined in (1a). Excise the adnominal adjunct to the head NP in NP$_1$. (It makes no difference in which direction a scan is made, and it can as well be back and forth):[7]

Noun suffixes occurring in the example sentences are as follows (where a suffix has more than one form, the first allomorph indicated is post-consonantal and the second one is post-vocalic): topic (TOP) *-ŭn/nŭn*; subject (SUBJ) *-i/ka*; object (OBJ) *-ŭl/lŭl*; genitive (GEN) *-ŭi* (ordinarily pronounced [e]); locative (LOC) *-e*; animate locative (LOC) *-eke*; person locative (LOC) *-tŏlŏ*; instrumental (INSTR) *-ŭlo/lo*; instrumental (INSTR) *-ŭlosŏ/losŏ*. Some example sentences also have occurrences of another class of elements called particles (PRTCLE).

Verb suffixes occurring in the example sentences are: literary declarative sentence-concluding (s.c.) *-ta* ; documentary declarative sentence-concluding (s.s.) *-ŭm/m*; past tense *-ŏs-/ŏt-*; conjunctive (CONJ) ko, *-ŏ*, *-ŏsŏ*, *-chimanŭn*, *-taka*, *-ŭmŭlo*.

Adnominal (ADN) suffixes are added to verb stems: *-ŭn/n*; *-nŭn*, *-ŭl/l*; *-tŏn* ; these differ in tense, and *-nŭn* occurs only with stems of action verbs. An adnominalized (ADN) verb in a sentence takes along with it all its complements, if it has any; it is essentially an adnominalized predicate. An adnominal phrase is an adnominalized predicate plus a noun head, to which the adnominalized predicate is adjunct and which it modifies. Adnominalized predicates are roughly equivalent to English relative clauses, while single adnominalized verbs are often equivalent to English adjectives (in the case of adjective verbs) and *-ed* and *-ing* forms (in the case of action verbs) functioning as adjectives.

6. In the strings of word classes, a hyphen separates the stem of a noun or verb from a suffix or particle.

7. The English equivalents given in the scanning steps are more literal, while the example sentences themselves are rendered in more idiomatic English.

ilŭl kanŭn uli ttal → *uli ttal*
my daughter, who is replacing her teeth → my daughter

Sentence (1) is now reduced to

(1) b. *nanŭn uli ttal unnŭn ŏlkulŭl chowahanta*
I like the smiling face of my daughter

(ii) Excise the adnominal adjunct to the head N in NP_2:

unnŭn ŏlkulŭl → *ŏlkulŭl*
smiling face → face

Sentence (1) is further reduced to the following, which shows a common sentence type :

(1) c. $(nanŭn)_1$ + $(uli\ ttal\ ŏlkulŭl)_2$ + $(chowahanta)_3$
$(\text{as for me})_1$ + $(\text{the face of my daughter})_2$ + $(\text{I like})_3$
$(\text{NP-TOP/SUBJ})_1$ + $(\text{NP-OBJ})_2$ + $(\text{Verb-s.c.})_3$

As to the assigning of the topic noun *na* "I" to the verb *chowahanta* "(one) likes", there is a general rule that the subject of a verb is not repeated when it is identical with the topic N or NP. In the case of sentence (1), no subject marked with the subject suffix *-i/ka* is specified, so it is assumed that it is the subject implied by the topic N or NP.

At this point, adjuncts that were excised in steps (i) and (ii) may be examined for further excision of adjunct material. In (1c) the NP *uli ttal ŏlkulŭl* "my daughter's face" contains the NP *uli ttal* "my daughter" as adjunct to *ŏlkulŭl* "face".

(iii) Excise the NP *uli ttal*:

uli ttal ŏlkulŭl → *ŏlkulŭl*
my daughter's face → face

Step (iii) further reduces sentence (1) while preserving the structure shown in (1c):

(1) d. *nanŭn ŏlkulŭl chowahanta.*
I like faces.

2.2 Sentences containing one or more conjunctive clauses

The structure of conjunctive clauses differs from that of sentence-concluding clauses, or independent sentences, only in that the suffix on the verb is different.

(2) *nanŭn kolmokŭl china kal ttae'e palŭl mŏmch'uko hanch'am sŏikkehanŭn p'iano solilŭl chowahanta.*[8]
"I like the sound of a piano that makes me stop and linger a while as I pass through a lane."

Re-write the words in (2) as a string of word classes:

(2) a. *na-nŭn kolmok-ŭl china kal ttae-e palŭl*
I.as.for lane at.the.time. I.pass.through feet
N-TOP N-OBJ ADN NP-LOC N-OBJ
mŏmch'u-ko hanch'am sŏikkeha-nŭn p'iano soli-lŭl
bring.to.a.halt.and for.a.while which.makes.me.linger piano sound
V-CONJ ADV V-ADN N N
chowaha-nta
like
V-S.C.

With the first scan, mark adjuncts for excision (here underlined).

(i) Excise the locative NP adjunct to the VP-CONJ *palŭl mŏmch'uko* "I bring my feet to a halt":
kolmokŭl china kal ttae'e palŭl mŏmch'uko → *palŭl mŏmch'uko*
at the time I pass through a lane I bring my feet to a halt → I bring my feet to a halt

(ii) Excise the adverb adjunct *hanch'am* to the adnominalized verb *sŏikke-hanŭn* "which makes me linger":

hanch'am sŏikkehanŭn → *sŏikkehanŭn*
which makes me linger a while → which makes me linger

Sentence (2) is now reduced to

(2) b. *nanŭn palŭl mŏmch'uko sŏikkehanŭn p'iano solilŭl chowahanta*

---

8. Because the sequence /ae/ represents a single phoneme, an apostrophe separates /ae/ and /e/ where the sequence represents the two phonemes /ae/ and /e/; thus: /*ttae'e*/ "at the time", where the noun /*ttae*/ "time" has added to it the locative suffix /*-e*/. In strings of word classes, as in (2a), where suffixes are separated from stems by a hyphen, the word /*ttae'e*/ appears as /*ttae-e*/. Similarly, in a case where the sequence /a/ and /e/ represents two phonemes, an apostrophe would separate the two symbols in order to distinguish the sequence from the unit phoneme /ae/; thus, for example, /*kilkka'e*/ or /*kilkka-e*/ "on the side of the road".

Examining now the segment *palŭl mŏmch'uko sŏikkehanŭn*, we ask whether the VP-CONJ *palŭl mŏmch'uko* "I bring my feet to a halt and" is an independent conjunctive clause or is a proper part of the VP that follows it (though the latter is an adnominalized form) *sŏikkehanŭn* "which makes me linger". One of the major functions of the conjunctive suffix *-ko* is coordinative "and", and the scope of a coordinative conjunctive clause in *-ko* can be ambiguous. If the segment *palŭl mŏmch'uko* is taken as a major constituent, as indeed one student did, the sentence must be interpreted as meaning "I stop when I go through a lane and I like the sound of a piano". This is less likely than taking it as a proper part of the following (adnominalized) VP *sŏikkehanŭn*, so that the segment *palŭl mŏmch'uko sŏikkehanŭn* would mean "which brings my feet to a halt and makes me linger". The basic structure of (2) is thus seen to be not

(2) c. *(*nanŭn*)$_1$ + (*palŭl mŏmch'uko*)$_2$+ (*sŏikkehanŭn p'iano solilŭl chowahanta*)$_3$
(N-TOP/SUBJ)$_1$ + (NP-OBJ)$_2$ + (V-s.c.)$_3$
(as for me)$_1$ (I bring my feet to a halt and)$_2$ (I like the sound of a piano which makes me linger)$_3$

but rather

(2) d. (*nanŭn*)$_1$ + (*palŭl mŏmch'uko sŏikkehanŭn p'iano solilŭl*)$_2$ + (*chowahanta*)$_3$
(N-TOP/SUBJ)$_1$ + (NP-OBJ)$_2$ + (V-s.c.)$_3$
(as for me)$_1$ (I like)$_3$ (the sound of a piano which brings my feet to a halt and makes me linger)$_2$

By going through the steps of string analysis, the student might have been alerted to the inherent technical ambiguity of the scope of the conjunctive clause in *-ko* in this case and might have avoided coming to a hasty conclusion as to the structure of the sentence.

3. Determining constituent segments and their relations

The methods of string analysis are especially helpful to the student when he meets long sentences whose basic structure seems to be hidden by the underbrush of adjunct material. As mentioned earlier, written sentences are made longer by conjunctive clauses, long and often complex adnominal adjuncts to head nouns, or both. The student may experience difficulty in two respects:

determining the scope of a conjunctive clause, that is, whether, as noted above, it stands as a major constituent of the sentence or is included as a proper part of a longer clause; and identifying the subject of a predicate, particularly when there are two or more predicates in the sentence. Often the problems of scope and of the subject–predicate tie are interrelated, so that by getting the one right the student will get the other right as well.

3.1 Sentences made longer by conjunctive clauses

Example (3) below shows how string analysis can help the student to determine the scope of a conjunctive clause and, consequently, recognize the correct tie between subject and predicate:

(3) *halmŏmŭi mŏnghi ttŭkoinnŭn nunenŭn kŭŭi adŭlkwa ttalkwa ch'ilsimnyŏn kan kosaengŭn hayŏtchimanŭn chŏngtŭn kohyangsanchŏni pich'inŭntŭthayŏtta.*
In the dreamy eyes of the old woman servant there seemed to be reflected her son and daughter and her home village of which she was so fond though she had lived a hard life there for seventy years.

Re-write the words of (3) as a string of word classes:

(3a) *halmŏm-ŭi mŏnghi ttŭkoin-nŭn nun-e-nŭn*
old.woman.servant-of dreamy which.are.open eyes-in-TOP
N-GEN ADV VP-ADN N-LOC TOP (NP$_1$)
kŭ-ŭi adŭl-kwa ttal-kwa ch'ilsimnyŏn kan kosaeng-ŭn
that.one-of son-and daughter-and 70.years space hardship-as.for
N-GEN N-PARTCLE N-PARTCLE N N N-TOP
ha-yŏtchimanŭn chŏngtŭ-n kohyangsanch'ŏn-i
she.did, but which.became.loved home.village
V-CONJ V-ADN N-SUBJ (NP$_2$)
pich'inŭntŭthayŏtta
seemed.to.be.reflected
V-S.C.

(i) A scan of (3a) shows a locative topic noun phrase NP$_1$, with a VP-ADN adjunct to the head noun of the NP, *nun* "eyes". Excise the VP-ADN adjunct:

halmŏmŭi mŏnghi ttŭkoinnŭn nunenŭn → halmŏmŭi nunenŭn
in the old woman servant's eyes which are dreamily open → in the eyes of the old woman servant

(ii) NP_2, which is marked as a subject by the subject suffix *-i/ka*, also has a VP-ADN adjunct.

Excise the VP-ADN adjunct:

> *kŭŭi adŭlkwa ttalkwa ch'ilsimnyŏn kan kosaengŭn hayŏtchimanŭn chŏngtŭn kohyangsanch'ŏni* → *adŭlkwa ttalkwa kohyangsanch'ŏni*
> her son and daughter and her home village of which she was so fond though she had lived a hard life there for 70 years → her son and daughter and her home village

Sentence (3) is now reduced to

(3) b. *halmŏmŭi nunenŭn adŭlkwa ttalkwa kohyangsanchŏni pich'inŭntŭthayŏtta*
There seemed to be reflected in the eyes of the old woman servant her son and her daughter and her home village

As (ii) shows, the segment *adŭlkwa ttalkwa* "son and daughter and" is grammatically coordinate with the segment *kohyangsanch'ŏn* "home village", even though a VP-CONJ, "of which she was so fond though she had lived a hard life there for 70 years", intervenes.

Adjunct material that has been excised in steps (i) and (ii) above may now be examined.

(iii) From the VP-ADN which was excised in step (i) above, excise the adverb adjunct *mŏnghi* "dreamily" to the adnominal verb *ttŭkoinnŭn* "which are open":

> *mŏnghi ttŭkoinnŭn* → *ttŭkoinnŭn*
> which were dreamily open → which were open

The VP-ADN which was excised in step (ii) above consists of two clauses: a conjunctive clause *ch'ilsimnyŏn kan kosaengŭn hayŏtchimanŭn* "though she had lived a hard life for 70 years" and a concluding clause *chŏngtŭn* "of which she was fond" in adnominalized form. The adnominal feature of *chŏngtŭn* extends to the conjunctive clause preceding it; this is why the VP-ADN in step (ii) translates as "of which she was so fond though she had lived a hard life there for 70 years".

(iv) Excise the conjunctive clause, which may be considered an adjunct to the concluding verb of the VP-ADN:

ch'ilsimnyŏn kan kosaengŭn hayŏtchimanŭn chŏngtŭn → *chŏngtŭn*
of which she was so fond though she had lived a hard life there for 70 years → of which she was so fond

Sentence (3) is now reduced to its essential structure:

(3) c. (*ttŭkoinnŭn nunenŭn*)$_1$ + (*adŭlkwa ttalkwa kohyangsanch'ŏni*)$_2$ + (*pich'inŭntŭthayŏtta*)$_3$
(NP-LOC TOP)$_1$ + (NP-SUBJ)$_2$ + (Verb-s.c.)$_3$
(As for in her dreamy eyes)$_1$ + (there seemed to be reflected)$_3$ + (her son and daughter and home village)$_2$.

Carrying out the procedures of string analysis for the purpose of formulating a grammar of the language would not stop at this point. For instance, the verb form at the end of the sentence, *pich'inŭntŭthayŏtta* "seemed to be reflected", is a complex verb form that can be analyzed. But for the purpose of the present demonstration it is sufficient to treat it as a unit verb form, because the student will recognize its morphological makeup and understand it as a semantic unit; analyzing it will not advance the step-by-step separation of the essential structure of the sentence from its adjunct material. The aim of this application of string analysis to Korean sentences, as stated in the Introduction, is to strip away only enough of the underbrush of adjunct material to reveal to the student the syntactic structure of the sentence.

Example sentence (3) above serves to illustrate once more the point that going through the steps of string analysis alone can help a student avoid various pitfalls in looking for the structure of a sentence. Assignment of subject–predicate ties in Korean sentences is often difficult; there are various structural circumstances that give rise to this difficulty. One of them is the case where the subject noun is present in the sentence but is not so marked by presence of the subject suffix. Students are likely to be wary of this possibility. But in this example, one student let himself be misled and assigned subject status to a noun phrase with no subject suffix while missing the real subject noun phrase even though it is marked by the subject suffix. The student took as the subject of the main verb phrase of the sentence *pich'inŭntŭthayŏtta* "seemed to be reflected" the segment *ch'ilsimnyŏn kan* "the space of seventy years" (which has no subject suffix) rather than the segment *kŭŭi adŭlkwa ttalkwa ch'ilsimnyŏn kan kosaengŭn hayŏtchimanŭn chŏngtŭn kohyang sanch'ŏni* "her son and daughter and her home village of which she was so fond though she had lived a hard life there for seventy years" (which has the subject suffix).

This is what led him to misconstrue the structure of example sentence (3) and conclude too quickly that the sentence meant *"In the old woman servant's dull eyes the seventy years she had toiled in the countryside and her son and daughter were reflected".

Judging from what this student left out in his translation, he also seems to be have been discouraged as well as misled by the sheer length of the adnominal phrase whose head is the noun phrase *kohyang sanch'ŏn-i* "home village-SUBJ". Another point in example sentence (3) which evidently confused the student was that he failed to realize that, while the segment *ch'ilsimnyŏn kan* "space of seventy years" is indeed a noun phrase, it has underlyingly the locative suffix *-e* "in", which has been (optionally) deleted, so that the noun phrase means "in the space of seventy years" or "for seventy years". As a systematic approach to perceiving the structure of sentences like example (3), applying the steps of string analysis would have guided the student to a correct reading of it by keeping him on the alert for alternative conclusions as to the structure of the sentence.

3.2 Sentences made longer by adnominal adjuncts

(4) *osipselŭl hwŏlssin nŏmŏssŭmeto pulkuhako chŏlmŏssŭl ttaeŭi mika achikto chiwŏchichiank'o namainnŭn tokil yŏchalosŏnŭn chinach'ike sŏmsehan sŏnŭi ŏlkule ipsulŭn kutke chamkyŏchyŏissŏkko achu nŏlpŭn hŭin ima'enŭn mwŏnchi molŭl kŭnŭli tŏphyŏinnŭnkŏtkat'atta.*
On her face, where the beauty of her youth had not yet faded but remained in spite of her being well over 50, and with features exceedingly fine for a German woman, her lips were tightly closed, and her very broad, pale forehead seemed to be covered with something of a shadow.

Re-write (4) as a string of word classes:

(4) a.

osipse-lŭl	*hwŏlssin*	*nŏmŏssŭm-e-to*	*pulkuhako*	
years.of.age	by.very.much	being.over-in-even	in.spite.of	
N-OBJ	**ADV**	**N-LOC-PRTCLE**	**ADV**	
chŏlmŏssŭlttae-ŭI		*mi-ka*	*achik-to*	*chiwŏchichiank'o*
time.when.she.was.young-of		beauty	yet-even	was.not.erased.and
N-GEN		**N-SUBJ**	**Adv.**	**V-CONJ**
namain-nŭn	*tokil*	*yŏcha-losŏ-nŭn*	*chinach'ike*	
which.remains	Germany	woman-for-as.for	exceedingly	
V-ADN	N	N-instr.-TOP	ADV	

sŏmseha-n *sŏn-ŭi* *ŏlkul-e* *ipsul-ŭn* *kutke*
which.are.delicate lines-of face-on lip-as.for tightly
V-ADN **N-GEN** **N-LOC** **(NP_1)** **N-TOP** **ADV**
chamkyŏchyŏissŏk-ko *achu* *nŏlp-ŭn* *hŭi–n*
were closed and very which.is.broad which.is.pale
V-CONJ **ADV** **V-ADN** **V-ADN**
ima-e-nŭn *mwŏnchi* *mol-ŭl*
forehead-on-as.for which.one.does.not.know.what.it.is
N-LOC TOP **(NP_2)** **VP-ADN**
kŭnŭl-i *tŏphyŏinnŭnkŏtkat'atta*
shade seemed to be put on
N-SUBJ **(NP_3)** **V-S.C.**

(i) Scanning (4a) above from left to right, excise the genitive adjunct to the head N *ŏlkule* "on her face" of locative NP_1:

osipselŭl hwŏlssin nŏmŏssŭmeto pulkuhako chŏlmŏssŭl ttaeŭi mika achikto chiwŏchichiank'o namainnŭn tokil yŏchalosŏnŭn chinach'ike sŏmsehan sŏnŭi ŏlkule → *ŏlkule*
on her face, where the beauty of her youth had not yet faded but remained in spite of her being well over 50 → on her face

This step reduces sentence (4) to:

(4) b. *ŏlkule ipsulŭn kutke chamkyŏchyŏissŏkko achu nŏlpŭn hŭin ima'enŭn mwŏnchi molŭl kunŭli tŏphyŏinnŭnkŏtkat'atta*
On her face her lips were tightly closed, and her broad pale forehead seemed to be covered with something of a shadow

(ii) Excise the adverb adjunct *kutke* "tightly" to the verb *chamkyŏissŏkko* "was closed and":

ŏlkule ipsulŭn kutke chamkyŏchyŏissŏkko → *ŏlkule ipsulŭn chamkyŏchyŏissŏkko*
on her face her lips were tightly closed and → on her face her lips were closed and

(iii) Excise the adnominal adjuncts, including the adverb tied to them, to the head N *ima'enŭn* "as for on her forehead" of NP_2:

achu nŏlpŭn hŭin imaenŭn → *imaenŭn*
as for on her very broad, pale forehead → as for on her forehead

This step reduces sentence (4) further to (4c) below:

(4) c. *ŏlkule ipsulŭn chamkyŏchyŏissŏkko ima'enŭn mwŏnchi molŭl kŭnŭli tŏphyŏinnŭnkŏtkat'atta*
On her face her lips were closed and her forehead seemed to be covered with something of a shadow

(iv) Excise the VP-ADN adjunct to the head N of NP_3, the subject N *kŭnŭl-i* "shade-SUBJ"

mwŏnchi molŭl kŭnŭli → *kŭnŭli*
something of a shadow → shadow

By this step, sentence (4) is now easily seen to consist of a frequently occurring sentence type:

(4) d. (*ŏlkule ipsulŭn chamkyŏchyŏissŏkko*)$_1$ + (*ima'enŭn kŭnŭli tŏphyŏ innŭnkŏtkat'atta*)$_2$
(Conjunctive clause)$_1$ + **(Concluding clause)**$_2$
(On her face her lips were closed and)$_1$ + (her forehead seemed to be covered with something of a shadow)$_2$

In Clause 1 the subject of the verb of the clause, *chamkyŏchyŏissŏkko* "were closed and", is not specified but is implied by the N-TOP, *ipsul* "lips". In Clause 2 the predicate of the clause, *tŏphyŏinnŭnkŏtkat'atta* "seemed to be covered by", is a complex verb form, and, like the verb form *pich'inŭntŭ-thayŏtta* "seemed to be reflected" in (3c), is best considered for the present purpose to be a unit verb form which ties to a single subject — in this case, *kŭnŭl-i* "shade", which has the subject suffix *-i/ka*.

The long adjunct that was excised in step (i) may now be examined if its structure is not clear. This segment is now easier to analyze once it has been isolated by excision.

(v) From the long NP-GEN adjunct to the locative N *ŏlkule* "on her face" in NP_1, which was excised in (i), excise the NP-GEN adjunct to the subject N *mi-ka* "beauty-SUBJ":

osipselŭl hwŏlssin nŏmŏssŭmeto pulkuhako chŏlmŏssŭl ttaeŭi mika → *mika*
the beauty of her youth in spite of her being well over 50 → her beauty

The NP-GEN adjunct excised in (v) may be analyzed in further detail if its structure is not clear. For instance, the segment *osipselŭl hwŏlssin nŏmŏssŭmeto*

pulkuhako "in spite of her being well over 50" will be recognized as an instance of a verbal pattern N-*e-to pulkuhako* "in spite of N", which functions as an adverbal expression; in this case it is adjunct to the adnominal phrase *chŏlmŏssŭl ttaeŭi* "of the time when she was young".

(vi) Excise the adverbal expression in (v):

> *osipselŭl hwŏlssin nŏmŏssŭmeto pulkuhako chŏlmŏssŭl ttaeŭi → chŏlmŏssŭl ttaeŭi*
> of the time when she was young though she was well over 50 → of the time when she was young

Looking now at the remaining segment of the NP-GEN that was excised in step (i), *mika achikto chiwŏchichiank'o namainnŭn tokil yŏchalosŏnŭn chinach'ike sŏmsehan sŏnŭi ŏlkule* "on her face with features exceedingly fine for a German woman, whose beauty had not yet faded but remained in spite of her being well over 50", it becomes clear that the subject N *mi-ka* "beauty" is the subject of the immediately following adnominal adjunct to the NP *tokil yŏchalosŏnŭn* "for a German woman".

(vii) Excise the adnominal adjunct to the NP *tokil yŏchalosŏnŭn* "as for for a German woman":

> *mika achikto chiwŏchichiank'o$_1$ namainnŭn$_2$ tokil yŏchalosŏnŭn chinach'ike sŏmsehan sŏnŭi ŏlkule → tokil yŏchalosŏnŭn chinach'ike sŏmsehan sŏnŭi ŏlkule*
> on her face with features exceedingly fine for a German woman, whose beauty had not yet faded but remained → on her face with features exceedingly fine for a German woman

The adnominal adjunct excised in step (vii) consists of two clauses, a coordinating conjunctive clause$_1$ and a concluding clause$_2$ (in adnominalized form), sharing a single subject *mi* "beauty".

(viii) Excise the conjunctive clause, which may be considered adjunct to the concluding verb or clause:

> *mika achikto chiwŏchichi ank'o namainnŭn → mika namainnŭn*
> her beauty had not yet faded but remained → her beauty remained

(ix) Finally, the NP-instr. *tokil yŏchalosŏnŭn* "as for for a German woman" and the adverb *chinach'ike* "exceedingly", which is adjunct to the adnominal

sŏmsehan "fine", may be excised. However, the adnominal *sŏmsehan* "fine" adjunct to the N-GEN *sŏnŭi* "of lines" should not be excised, as in this instance *sŏn* cannot stand semantically independent of a modifying adjunct.

tokil yŏchalosŏnŭn chinach'ike sŏmsehan sŏnŭi ŏlkule → *sŏmsehan sŏnŭi ŏlkule*
on her face with features exceedingly fine for a German woman → on her face with fine features

The treatment of sentence (4) shows how the progressive excising of adjunct material in steps from larger to smaller segments first isolates the basic structure of the sentence as a whole and then isolates the structure of the longer excised segments down to any needed level of detail.

3.3 Sentences made longer by a series of conjunctive clauses

Here the purpose of the string analysis will be to bring out the structure of a long sentence while preserving the identity of its conjunctive clause constituents. Sentence (5) is a standard degree-conferring certificate, an example of documentary style.

(5) *wi salamŭn ponkyo taehakwŏn sŏksa kwachŏngŭl isuhako sochŏngŭi sihŏme hapkyŏkhayŏ chech'ulhan araeŭi nonmuni simsa'e t'ongkwatweŏ ihak sŏksaŭi chakyŏkŭl katch'uŏssŭmŭlo ilŭl inchŏngham.*
[This certificate] confirms that the above-named person, his thesis named below, which he submitted upon completing the course of study for the Master's degree in the Graduate School of this University and passing the required examinations, having been accepted, has satisfied the requirements for the degree of Master of Science.

Re-write (5) as a string of word classes:

(5) a.

wi	*salam-ŭn*	*ponkyo*	*taehakwŏn*		*sŏksa*
above	person-as.for	this.school	Graduate.School		Master
N	N-TOP	N	N		N
kwachŏng-ŭl isuha-ko			*sochŏng-ŭi*		*sihŏm–e*
course	upon.completing-and		requirement-of		examination-in
N-OBJ	$V\text{-CONJ}_1$		N-GEN		N-LOC
hapkyŏkha-yŏ	*chech'ulha-n*		*arae-ŭi*	*nonmun-i*	*simsa-e*
upon.passing	which.he.submitted		below-of	thesis	judging-in
$V\text{-CONJ}_2$	ADNOM V		N-GEN	N-SUBJ	N-LOC

t'ongkwatwe-ŏ *ihak* *sŏksa-ŭi* *chakyŏk-ŭl* *katch'uŏssŭmŭlo*
having-been.passed science Master-of qualification he.has-because
V-CONJ$_3$ **N** **N-GEN** **N-OBJ** **V-CONJ$_4$**
i-lŭl *inchŏngham*
this [this certificate] confirms
N-OBJ **V-S.C.**

Scanning (5a), it can be seen that it contains four conjunctive clauses plus a concluding clause. V-CONJ$_4$ is the end point of a long conjunctive verb phrase that stretches back to the beginning of the sentence. The remaining segment *ilŭl inchŏngham* "[this certificate] confirms this" is the concluding clause to V-CONJ$_4$ and is a simple sentence of the form (N-OBJ) + (V); the subject is implied from context rather than specified. The segment contains no adjunct material, and it can be put aside for the moment.

(i) Now scanning VP-CONJ$_4$ from right to left, separate off the introductory NP-TOP *wi salamŭn* "as for the above-named person". Identify conjunctive clause breaks within VP-CONJ$_4$; there are three, as noted below:

> *wi salamŭn* + (*ponkyo taehakwŏn sŏksa kwachŏngŭl isuhako$_1$ sochŏngŭi sihŏme hapkyŏkhayŏ$_2$ chech'ulhan araeŭi nonmuni simsa'e t'ongkwatweŏ$_3$)$_{segment\ 1}$* + (*ihak sŏksaŭi chakyŏkŭl katch'uŏssŭmŭlo$_4$)$_{segment\ 2}$*
> as for the above-named person + (his thesis named below, which he submitted upon completing the course of study for the Master's degree in the Graduate School of this University$_1$ and passing the required examinations$_2$, having been accepted$_3$)$_{segment\ 1}$ + (because he has the qualifications for the degree of Master of Science$_4$)$_{segment\ 2}$

Segment 2 of (i) is of the simple sentence type, NP-OBJ + V (the subject of V here is "he", implied by the sentence-introducing NP-TOP *wi salamŭn* "as for the above-named person").

Segment 1 of (i) consists of the segments NP-SUBJ + VP-CONJ$_3$. The VP is simple, consisting only of N-LOC + V, *simsa'e t'ongkwatweŏ* "was accepted", and containing no adjunct material. The NP-SUBJ, on the other hand, is rather complex and contains adjunct material to the N-SUBJ *nonmun-i* "thesis-SUBJ"
(ii) Excise the adnominal adjunct to the N head of the adnominal NP. Notice that the adnominal adjunct itself consists of VP-CONJ$_3$ + a concluding clause in adnominal form. The NP-TOP introducing sentence (5), *wi salamŭn* "the above person", is not excised because it is not an adjunct:

wi salamŭn ponkyo taehakwŏn sŏksa kwachŏngŭl isuhako sochŏngŭi sihŏme hapkyŏkhayŏ chech'ulhan araeŭi nonmuni → *wi salamŭn araeŭi nonmuni*
as for the above-named person, his thesis named below which he submitted, upon completing the Master's course in the Graduate School of this University and passing the required examinations → as for the above-named person, his thesis named below

The VP-conj$_4$ of sentence (5), now reduced to (5b) below, is seen to consist of VP-CONJ$_3$ + a concluding clause:

(5) b. (*wi salamŭn*) + (*araeŭi nonmuni simsa'e t'ongkwatweŏ*)$_3$ + (*ihak sŏksa chakyŏkŭl katch'uŏssŭmŭlo*)$_4$
(as for the above-named person) + (his thesis named below having been passed)$_3$ + (because he has the qualifications for the Master of Science degree)$_4$

(iii) Excise VP-CONJ$_3$:

wi salamŭn araeŭi nonmuni simsa'e t'ongkwatweŏ ihak sŏksa chakyŏkŭl katch'uŏtssŭmŭlo → *wi salamŭn ihak sŏksa chakyŏkŭl katch'uŏssŭmŭlo*
because the above-named person, his thesis below having been passed, has the qualifications for the Master of Science degree → because the above-named person has the qualifications for the Master of Science degree

Putting (iii) together with the segment *ilŭl inchŏngham* of (5a) above that was set aside momentarily, sentence (5) is now reduced to

(5) c. *wi salamŭn ihak chakyŏkŭl katch'uŏssŭmŭlo ilŭl inchŏngham.*
Because the above-named person has the qualifications for the Master of Science degree, [this certificate] confirms this.

Each excised conjunctive clause can now be analyzed in greater detail if need be, excising the adjunct material in each clause in turn:

(iv) Excise from VP-CONJ$_1$ the noun *ponkyo* "this school", adjunct to the noun *taehakwŏn* "graduate school"; then excise *taehakwŏn* , which is adjunct to the NP *sŏksa kwachŏng* "Master's course"; these two steps can be taken as one:

wi salamŭn ponkyo taehakwŏn sŏksa kwachŏngŭl isuhako → *wi salamŭn sŏksa kwachŏngŭl isuhako*
the above-named person has completed the Master's course of the Graduate School of this University and → the above-named person has completed the Master's course

(v) Excise from VP-CONJ$_2$ the N-GEN *sochŏngŭi* "required", adjunct to *sihŏm* "examination":

sochŏngŭi sihŏme hapkyŏkhayŏ → *sihŏme hapkyŏkhayŏ*
having passed the required examinations → having passed the examinations

(vi) Excise from VP-CONJ$_3$ the N-GEN *araeŭi* "of below", adjunct to the N *nonmun-i* "thesis-SUBJ":

araeŭi nonmuni → *nonmuni*
the thesis below → the thesis

The basic structure of conjunctive clause$_4$ in (5) can now be seen as follows, highlighted in bold type. Conjunctive clause$_4$ relates to the three conjunctive clauses within it as the concluding clause to conjuctive clause$_3$, which in turn includes conjunctive clause$_1$ and conjuctive clause$_2$:

(5) d. (***wi salamŭn***)$_{\text{TOP/SUBJ}}$ *ponkyo taehakwŏn sŏksa* ((((***kwachŏngŭl isuhako*** $_{\text{V-CONJ1}}$) (*sochŏngŭi* ***sihŏme hapkyŏkhayŏ*** $_{\text{V-CONJ2}}$)) (***chech'ulhan*** *araeŭi* ***nonmuni simsa'e t'ongkwatweŏ*** $_{\text{V-CONJ-3}}$)) (*ihak sŏksaŭi* ***chakyŏkŭl katch'uŏssŭmŭlo*** $_{\text{V-CONJ4}}$)) + (*ilŭl inchŏngham* $_{\text{Concluding clause 5}}$).
(**The above-named person**,) ((((**his thesis** named below, **which he submitted upon completing the course of study** for the Master's degree in the Graduate School of this University **and**$_1$) (**passing the** required **examinations**$_2$,)) **having been accepted**$_3$,)) (**because he has satisfied the requirements for the degree of Master of Science**$_4$,)) + ([this certificate] confirms this$_5$.)

Or, schematically:

((((V-CONJ$_1$) (V-CONJ$_2$)) (V-CONJ$_3$)) (V-CONJ$_4$)) + (Concluding clause$_5$)

It can be inferred from the topic *wi salamŭn* "as for the above person" that the pronominal subject is shared by both VP-CONJ$_1$ and VP-CONJ$_2$, since no subject noun is specified and VP-CONJ$_1$ is coordinate with VP-CONJ$_2$. VP-CONJ$_3$, however, does have a specified subject, *nonmun-i* "thesis-SUBJ", the subject of *t'ongkwatweŏ* "having been passed". No subject of the verb in the concluding clause *ilŭl inchŏngham* is specified, nor does the topic *wi salamŭn* "as for the above person" imply a pronominal subject for it. This is a typical case where the pronominal subject of a verb must be inferred from the context, verbal or situational; here, perhaps "this (certificate)" or "we".

4. Benefits of string analysis of shorter sentences

It has been shown that applying the method of string analysis may help the student perceive the basic structure of long sentences. That a sentence is short, however, does not guarantee that its basic structure will be correctly perceived at once. This section demonstrates the usefulness of string analysis with regard to identifying even in shorter sentences the scope and structure of adnominalized predicates and the subject–predicate tie.

4.1 Recognizing the scope and structure of adnominalized predicates

(6) *salkie sitallyŏsŏ yangpankwa kwalli ppun anila chŏngpuwa kakkaun wekuk selyŏk kkachito paechŏkhaketwen nongmintŭleke tonghakŭi kalŭch'imŭn pankaŭn maliŏtta.*
The teachings of Eastern Learning were welcome words to the peasants, who, because of the bitterness of their lives, had come to reject not only the nobility and officials but even the foreign forces close to the government.

Re-write (6) as a string of word classes:

(6a) *salki-e sitally-ŏsŏ yangpan-kwa kwalli ppun*
life-in were.badly.treated-because nobility-and officials only
N-LOC V-CONJ N-PRTCLE N N
anila chŏngpu-wa kakkau-n wekuk
is.not.but government-and who.were.close foreign.country
V-CONJ N-PRTCLE V-ADNOM N
selyŏk-kkachi-to paechŏkhaketwe-n nongmin-tŭl-eke
forces-up.to-even who had come to reject peasant-pl.-to
N-PRTCLE-PRTCLE V-ADN N-PRTCLE-LOC(NP_1)
tonghak-ŭi kalŭch'im-ŭn pan'kau-n mal-iŏtta
Eastern Learning-of teachings-as for which were welcome words-were
N-GEN N-TOP V-ADNOM N-V-S.C.

(i) Excise NP_1, which is an adnominal phrase with a locative suffix *-eke* "to (a person)":

salkie sitallyŏsŏ yangpankwa kwalli ppun anila chŏngpuwa kakkaun wekuk selyŏk kkachito paechŏkhaketwen nongmintŭleke tonghakŭi kalŭch'imŭn pankaun maliŏtta → tonghakŭi kalŭch'imŭn pankaun maliŏtta

The teachings of Eastern Learning were welcome words to the peasants, who, because of the bitterness of their lives had come to reject not only the nobility and officials but even the foreign forces who were close to the government → The teachings of Eastern Learning were welcome words

It is immediately obvious that the adnominal phrase that has just been excised has the noun *nongmin* "peasants" for its head. The question at this point is, how far does the adnominalized predicate stretch?

(ii) Excise adjunct material from the adnominalized predicate in NP_1 step by step:

salkie sitallyŏsŏ yangpankwa kwalli ppun anila chŏngpuwa kakkaun wekuk selyŏk kkachito paechŏkhaketwen → salkie sitallyŏsŏ yangpankwa chŏngpuwa kakkaun wekuk sekyŏkkachito paech'ŏkhaketwen
who because of the bitterness of their lives had come to reject not only the nobility and officials but even the foreign forces close to the government → who because of the bitterness of their lives had come to reject even the foreign forces close to the government

(iii) *salkie sitallyŏsŏ chŏngpuwa kakkaun wekuk selyŏk kkachito paech'ŏkhaketwen → salkie sitallyŏsŏ wekuk selyŏkkachito paech'ŏkhaketwen*
who because of the bitterness of their lives had come to reject even the foreign forces close to the government → who because of the bitterness of their lives had come to reject even the foreign forces

Consider now the clausal structure of the remaining segment in (iii):

salki-e sitally-ŏsŏ wekuk selyŏk kkachito paech'ŏkhaketwe-n
N-LOC V-CONJ N N PARTCLE V-ADN

It can now be quickly seen that this simplified version of the long adnominalized predicate in NP_1 shows the frequently occurring structure (Conjunctive clause)$_1$ + (Concluding clause)$_2$, although Concluding clause$_2$ is in adnominalized form.

(iv) Excise the N-GEN *tonghakŭi* "of Eastern Learning", adjunct to the N *kalŭch'im* "teachings", in the segment isolated by the removal of NP_1 in step (i) above:

tonghakŭi kalŭch'imŭn pankaun maliŏtta → kalŭch'imŭn pankaun maliŏtta

the teachings of Eastern Learning were welcome words → the teachings were welcome words

The two segments of sentence (6) left after steps (iii) and (iv) above are now put back together again, replacing the N head of NP_1; this maintains the clausal structure of the sentence:

(6) a. (*salkie sitallyŏsŏ wekuk selyŏk kkachito paech'ŏkhaketwen nongmintŭleke*) + (*kalŭch'imŭn pankaun maliŏtta*)
(to the peasants who had come to reject even the foreign forces because of the bitterness of their lives) + (the teachings were welcome words)

The scope of the adnominal phrase NP_1 of sentence (6) is thus made clear, as shown in bold type:

(6) c. ***salkie sitallyŏsŏ yangpankwa kwalli ppun anila chŏngpuwa kakkaun wekuk selyŏk kkachito paechŏkhaketwen*** *nongmintŭleke kalŭch'imŭn pankaŭn maliŏtta.*
The teachings were welcome words to the peasants **who, because of the bitterness of their lives, had come to reject not only the nobility and officials but even the foreign forces close to the government.**

While sentence (6) above is itself relatively short, the adnominal phrase within it is quite long. The sheer length of an adnominal noun phrase can sometimes divert the reading eye away from its head noun even though the position of the head noun at the end of the phrase is rigidly fixed. Applying the procedures of string analysis to sentence (6) would surely have alerted one student to identify the head noun of the long adnominal phrase and thus also the adnominalized verb that is adjunct to it: *salkie sitallyŏsŏ yangpankwa kwalli ppun anila chŏngpuwa kakkaun wekuk selyŏk kkachito* ***paech'ŏkhaketwen nongmintŭleke*** "**to the peasants, who had come to reject** not only the nobility and officials but even the foreign forces because of the bitterness of their lives". The student evidently imagined, instead, a sentence structure for (6) which led to the translation *"To the farmers, for whom life was difficult, the teachings of Eastern Learning to resist not only the nobles and officials but also the government and foreign influence, these were welcome words." (Other errors by this student compounded the damage of the larger misreading of the structure of sentence (6) but they would be outside the purview of this chapter.)

Establishing the scope of adnominal phrases or of conjunctive clauses can be a delicate procedure. In making excisions of adjunct material we need to have as a result a sentence of the language, i.e., preserve "sentencehood".[9] We also need to preserve the basic meaning of the sentence that was originally intended. Consider a sentence like (7) below:

(7) *kŭliko toni ŏpsŏsŏ chŏktchohaechin ch'inkutŭlŭl uli chipe ch'ŏnghaepokosipta.*
And I want to invite my friends, with whom I've lost contact for lack of money.

Re-writing (8) as a string of word classes:

(7) a. *kŭliko ton-i ŏpsŏsŏ chŏktchohaechin*
and money is.lacking.,so with whom I have lost contact
Adv. N-subj V-conj V-adn
ch'inku-tŭl-ŭl uli chip-e ch'ŏnghaepokosipta
friend-s we house-to want to invite
N-prtcle-obj N N-loc V-s.c.

In sentence (7) the question is, what is the scope of the conjunctive clause *toni ŏpsŏsŏ* "money is lacking, so" or "because I have no money" or "for lack of money"? If we excise the adnominalized predicate *chŏktchohaechin* "with whom I've lost contact" as adjunct to *ch'inkutŭl* "friends", as one student did, we have as a result the sentence in (7b) below, with its major constituents being conjunctive clause + concluding clause after the introductory sentence adverb *kŭliko* "and":

(7) b. *(*kŭliko*) (*toni ŏpsŏsŏ*)$_1$ + (*ch'inkutŭlŭl uli chipe ch'ŏnghaepokosipta*)$_2$.
*And (because I have no money)$_1$ + (I want to invite my friends to my house)$_2$.

The sentence meaning obtained from the structure shown in (7b) strikes one as odd, since it is not likely that a person would ordinarily invite friends because he has no money. Yet, it is the only interpretation possible given the structure shown in (7b).

A second scan of (7) lets the adnominalized predicate *chŏktchohaechin* stand, so that the scope of the adnominal noun phrase *chŏktchohaechin ch'in-*

9. Harris (1962), p.23.

kutŭlŭl "friends with whom I've lost contact" would extend to the conjunctive clause before it, *toni ŏpsŏsŏ* "because I have no money". To put it another way, allowing the adnominalized predicate *chŏktchohaechin* to stand would show that the conjunctive clause relates not to the verb that comes last in the sentence, *ch'ŏnghaepokosipta* "I want to invite", but to the verb that immediately follows it, though in adnominalized form, *chŏktchohaechin* "with whom I've lost contact". And this is usually the case, that conjunctive clauses relate directly to the immediately following verb. We would then have the result of (7c) below, with the conjunctive clause a proper part of a long adnominal phrase *toni ŏpsŏsŏ chŏktchohaechin ch'inkutŭl* "friends with whom I've lost contact for lack of money". This structure is reflected in the English translation in (7) above, and repeated in (7c).

(7) c. (*kŭliko*) + (*toni ŏpsŏsŏ chŏktchohaechin ch'inkutŭlŭl*)$_2$ + (*uli chipe ch'ŏnghaepokosipta*)$_1$.
(And) + (I want to invite to our house my friends)$_1$ (with whom I've lost contact for lack of money.)$_2$

The sentence meaning obtained from the structure shown in (7c) no doubt makes much more cultural sense than that of (7b): you may have lost contact with friends if you don't have money to entertain them with.

4.2 Identifying the subject–predicate tie

Matching up the right one of several noun phrases marked as subjects by their suffixes with the right one of several predicates sometimes proves especially difficult when one of the subjects of one of the predicates is not specified by a noun word or phrase but is only implied. As has been suggested earlier, a number of factors, both verbal and situational, go into identifying the pronominal subject of a predicate when no subject noun is specified. String analysis is a useful first approach because it reduces the number of possibilities. Consider the following example:

(8) *kŭlonte halunŭn han tongnie sanŭn ch'we tŏkch'unilanŭn nongmini cho ung chŏnŭl sakochie pekkitaka natŏlŏ swinŭn t'ŭmt'ŭme ssŏtallanŭn put'akŭl hanŭnkŏsiŏtta.*
Now what happened was that one day a farmer called Ch'we Tŏkch'un who lived in the same village was copying out "The Tale of Cho Ung" on some writing paper when he asked me to do it for him in my leisure moments.

Re-write (8) as a string of word classes:

(8) a. *kŭlonte halu-nŭn han tongni-e sa-nŭn*
now one.day-as.for one village-in who.was.living
ADV N-TOP N N-LOC V-ADN
ch'we tŏkch'un ilanŭn nongmin-i cho ung chŏn-ŭl
Ch'we.Tŏkch'un who.is.called farmer Cho.Ung life.story
(proper noun) V-ADN N-SUBJ (proper noun) N-OBJ
sakochi-e pekki-taka na-tŏlŏ swi-nŭn
writing.paper-on was.copying.when me-to when.resting
N-LOC V-CONJ N-LOC V-ADNOM
t'ŭmt'ŭm-e ssŏtallanŭn put'ak-ŭl
intervals-in which.asked.me.to.write request
N-LOC V-ADNOM N-OBJ
hanŭnkŏs-ŏtta.
what.he.did.was.make
V-S.

Proceeding from left to right,

(i) Excise the introductory word *kŭlŏnte* "now (sentence-introducer)", which functions as an adverbal adjunct to the sentence as a whole.

(ii) Excise the N-TOP *halunŭn* "as for one day", which also functions as an adverbal adjunct to the sentence as a whole.

(iii) Excise the adnominal *han tongnie sanŭn* "who lives in the same village", which is adjunct to the NP *ch'we tŏkch'unilanŭn nongmin* "a farmer who is called Ch'we Tŏkch'un".

Sentence (8) is now reduced to

(8) b. *ch'we tŏkch'unilanŭn nongmini cho ung chŏnŭl sakochie pekkitaka natŏlŏ swinŭn t'ŭmt'ŭme ssŏ tallanŭn put'akŭl hanŭnkŏsiŏtta*
What happened was that a farmer by the name of *Ch'we Tŏkch'un* was copying "The Tale of Cho Ung" on some writing paper when he asked me to write it for him in my leisure moments.

(iv) Excise from (8b) the locative NP *swinŭn t'ŭmt'ŭme* "in the intervals when resting", leaving

(8) c. *ch'we tŏkch'unilanŭn nongmini cho ung chŏnŭl sakochie pekkitaka natŏlŏ ssŏ tallanŭn put'akŭl hanŭnkŏsiŏtta*
What happened was that a farmer by the name of Ch'we Tŏkch'un was copying "The Tale of Cho Ung" on some writing paper when he asked me to write it for him.

Sentence (8c) now shows one N marked as subject, *nongmin-i* "farmer-SUBJ", and this subject is easily assigned to the V-CONJ that immediately follows it, *pekkitaka* "was copying, when" as well as to the segment *ssŏ tallanŭn put'akŭl hanŭn kŏsiŏtta* "what he did was ask [me, the addressee] to write it for [him, the speaker]". The segment *ssŏ tallanŭn put'akŭl hanŭn kŏsiŏtta* "what he did was ask [me, the addressee] to write it for [him, the speaker]" is a complex form which, for the present purpose, may be treated as a unit V.

The question remaining is, What is the subject of the V-ADN *swinŭn* in the segment *swinŭn t'ŭmt'ŭme* "in the intervals when resting", which was excised in step (iv) above? The student is likely to see two possibilities: it could be the N *nongmin-i* "farmer-SUBJ" (or its pronominal replacement) or it could be some other N, which, however, is not specified in the sentence, for there is no other N marked as subject. That is, does the segment *swinŭn t'ŭmt'ŭme* mean "in the intervals while he (the farmer) is resting" or "in the intervals while I am resting"? Because the N *nongmin-i* has the subject suffix the student may jump to the conclusion that it is the subject of *swinŭn t'ŭmt'ŭme*. But the presence of the N *natŏlŏ* "to me" in the segment *natŏlŏ . . . put'akŭl hanŭnkŏsiŏtta* "what the farmer did was make a request to me . . ." implies that the subject of *swinŭn t'ŭmt'ŭme* is "I", i.e. "in the intervals when I am resting", or "in my leisure moments".

By directing his attention to the smaller segments set off by string analysis, the student is more likely to avoid a hasty conclusion, such as resulted in one student's mistranslation of example sentence (8) as "However, one day a farmer named Ch'we Tŏkch'un, who lived in the same village, while copying out "The Tale of Cho Ung", every time he would rest he would ask that I write it for him.' The student wrongly assigned the subject N *nongmini* "farmer-SUBJ" to the adnominal noun phrase *swinŭn t'ŭmt'ŭme* "in the intervals when resting" as well as to the verb *pekkitaka* "was copying, when" though correctly to *put'akŭl hanŭnkŏsiŏtta* "what he did was make a request". That the structure of the Korean sentence confused the student is evidenced also by the awkwardness of the English sentence.

5. Summary

This chapter has tried to show that the method of Z. S. Harris's string analysis is effective and efficient in helping students perceive the basic structure of written Korean sentences, especially longer ones. Perception of the basic structure is sometimes hindered by two main features of Korean sentences: long and complex adnominalized predicates, whose scope and head noun or noun phrase may be difficult for the student to pick out; and conjunctive clauses whose scope may not be immediately apparent to the student. The procedures of string anlysis lead to the isolation of the basic, or elementary, sentence or sentences undergirding the sentence as written. These procedures need be carried out down only to the level of detail at which the student can recognize the basic structure of the sentence.

Appendix: Table of English phonetic approximations

This table lists the phonemes, their main allophones and approximate phonetic equivalents in English. The symbols used for the phonemes are taken from the McCune-Reischauer system of transcription, a widely used conventional system. However, the symbols used for the allophones are taken from IPA (1999).

The consonant series /p t ch k s/ is lenis and slightly aspirated initially. The consonants /p t ch k/ are regularly voiced between voiced sounds. The consonant /s/ is regularly palatalized before the front vowel [i] and before the semivowel [j]; it may be slightly voiced after a nasal.

The consonant series /pp tt tch kk ss/ is fortis and unaspirated and with some glottal constriction; these consonants have but one allophone each and have a highly restricted distribution.

The consonant series /p' t' ch' k'/ is fortis and aspirated; these consonants also have but one allophone each and have a highly restricted distribution. (A more detailed description of Korean phonetics by Hyun Bok Lee may be found in IPA (1999), pp. 120–123.)

Consonants	Allophones	Eng. approx.	Allophones	Eng. Approx.
/p/	[p̥]	**p**ut, cu**p**	[b]	o**b**oe
/t/	[t̥]	**t**ake, pi**t**	[d]	i**d**ea
/ch/	[ʧ]	**ch**um	[ʤ]	le**g**ion
/k/	[g̊]	**k**arma, ra**ck**	[g]	le**g**al
/s/	[z̥]	**s**alad	[ʃ]	**sh**e
/pp/	[p]	s**p**y		
/tt/	[t]	s**t**op		
/tch/	[c]	righ**t ch**ange		
/kk/	[k]	s**k**ate		
/ss/	[s]	ma**ss s**urrender		
/p'/	[p^h]	a**pp**oint		
/t'/	[t^h]	pre**t**end		
/ch'/	[c^h]	a**ch**ieve		
/k'/	[k^h]	a**cc**ount		
/m/	[m]	**m**om		
/n/	[n]	**n**one		
/ng/	[ŋ]	sin**g**er, sin**g**		
/l/	[l]	**l**eak	[ɾ]	ci**t**y
/h/	[h]	**h**ot		
Vowels				
/i/	[i]	s**ea**t		
/e/	[e]	b**e**t		
/ae/	[æ]	h**a**t		
/a/	[ɑ]	s**o**lid		
/o/	[o]	**o**rder		
/ŏ/	[ʌ]	n**u**t		
/u/	[u]	m**oo**n		
/ŭ/	[ɯ]	b**oo**k		
Semi-vowels				
/y/	[j]	**y**es		
/w/	[w]	**w**e		

References

Harris, Zellig S. 1962. *String Analysis of Sentence Structure*. The Hague: Mouton.

IPA. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press

Lukoff, Fred. 1982. *An Introductory Course in Korean*. Seoul: Yonsei University Press.

Lukoff, Fred. 1985. "Perceiving the structure of long sentences in written Korean". In Kim, Nam-Kil & Henry H. Tiee, eds. (1985), *Studies in East Asian Linguistics*, pp. 192–209. Los Angeles: University of Southern California.

Zellig Sabbettai Harris

A comprehensive bibliography of his writings, 1932–2002*

Compiled by

E. F. K. Koerner
University of Ottawa

1932. *Origin of the Alphabet.* Unpublished M.A. thesis, University of Pennsylvania, Philadelphia, Pa., 111 typed pp.

1933. "Acrophony and Vowellessness in the Creation of the Alphabet". *Journal of American Oriental Society* 53:387. [Summary of 1932 thesis.]

1934a. "The Structure of Ras Shamra C". *Journal of the American Oriental Society* 54:80–83.

1934b. Review of Raymond P[hilip] Dougherty [(1877–1933)], *The Sealand of Ancient Arabia* (New Haven, Conn.: Yale University Press; London: Oxford University Press, 1932). *Journal of the American Oriental Society* 54:93–95.

1935a. Review of Edward Chiera, *Joint Expedition [of the American School of Oriental Research in Bagdad] with the Iraq Museum of Nuzi*, vols. 4–5 (Paris: P. Geuthner, 1933–1934). *Language* 11:262–263.

1935b. "A Hurrian Affricate or Sibilant in Ras Shamra". *Journal of the American Oriental Society* 55:95–100.

1936a. (Together with James A[lan] Montgomery [(1866–1949)].) *The Ras Shamra Mythological Texts.* (=*Memoirs of the American Philosophical Society*, 4.) Philadelphia: American Philosophical Society, 134 pp.
Reviewed by
Edward Sapir (1884–1939) in *Language* 13:326–331 (1937).

1936b. *A Grammar of the Phoenician Language.* (=*American Oriental Series*, 8.) New Haven, Conn.: American Oriental Society, xi, 172 pp.

---

* A first version of the present list appeared in *Historiographia Linguistica* 20:2/3.509–522 (1993), pp. 510–520. (The "Introductory remarks" [p. 509] and Appendix "Appraisals of Zellig S. Harris, 1962–1993" [520–522] are omitted here). I'd like to thank Bruce E. Nevin for his corrections and additions.

[Ph.D. dissertation, University of Pennsylvania, Philadelphia, 1934.]
Reviewed by
Edward Sapir in *Language* 15: 60–65 (1939);
Vojtěch Šanda (1873–*post* 1939) in *Archiv Orientální* 11: 177–178 (1939);
Charles François Jean (1874–1965) in *Revue des Études Sémitiques* 1940: 94–96;
Maria Höfner (1901–1995?) in *Wissenschaftliche Zeitschrift für die Kunde des Morgenlandes* 48: 153 (1941).

1936c. "Back Formation of *itn* in Phoenician and Ras Shamra". *Journal of the American Oriental Society* 56: 410. [Abstract.]

1937. "A Conditioned Sound Change in Ras Shamra". *Journal of the American Oriental Society* 57: 151–157.

1938a. "Expression of the Causative in Ugaritic". *Journal of the American Oriental Society* 58: 103–111.

1938b. "Ras Shamra: Canaanite civilization and language". *Annual Report of the Smithsonian Institution 1937*: 479–502; illus., 4 pl., 1 map on 2 leaves. Washington, D.C.

1939a. *Development of the Canaanite Dialects: An investigation in linguistic history*. (= *American Oriental Series*, 16.) New Haven, Conn.: American Oriental Society, x, 108 pp.; illus., map. (Repr., Millwood, N.Y.: Kraus, 1978.)
Reviewed by
William Foxwell Albright (1891–1971) in *Journal of the American Oriental Society* 60: 414–422 (1940);
René Dussaud (1868–1958) in *Syria: Revue d'art oriental et d'archéologie* 1940: 228–230 (Paris);
Gonzague Ryckmans (1887–1969) in *Le Muséon* No. 53: 135 (1940);
Harold Louis Ginsberg (b.1903) in *Journal of Biblical Literature* 59: 546–551 (1940);
Max Meir Bravmann (1909–1977?) in *Kirjath Sepher* 17: 370–381 (1940);
Marcel Cohen (1884–1974) in *Bulletin de la Société de Linguistique de Paris* No. 123: 62 (1940–1941);
Raphaël Savignac in *Vivre et Penser* 1941: 157–159;
Albrecht Goetze (1897–1971) in *Language* 17: 167–170 (1941);
Alexander Mackie Honeyman (1907–1988) in *Journal of the Royal Asiatic Soci-ety of Great Britain and Ireland* 17: 167–170 (1941);
Bernard Baron Carra de Vaux (1867–*c.*1950) in *Journal of the Palestine Oriental Society* 19: 329–330 (Jerusalem, 1941);
Franz Rosenthal (b.1914) in *Orientalia* 11: 179–185 (1942);
Ronald J[ames] Williams (b.1917) in *Journal of Near Eastern Studies* 1: 378–380 (Chicago, 1942).

1939b. "Development of the the West Semitic Aspect System". *Journal of the American Oriental Society* 59: 409–410. [Abstract.]

1939c. (Together with Charles F. Voegelin [(1906–1986)].) *Hidatsa Texts Collected by Robert H. Lowie*, with grammatical notes and phonograph transcriptions by Z. S. Harris & C. F. Voegelin. (= *Prehistory Research Materials*, 1:6), 173–239. Indianapolis: Indiana Historical Society. (Repr., New York: AMS Press, 1975.)
[According to Harris (1990, 2002), in the late 1930s and early 1940s he used this

Hidatsa material, together with the Kota texts that are reviewed in (1945f), for his development of substitution grammar, discourse analysis, and transformational analysis, research which was demonstrated at the Linguistic Institute and subsequently first published in (1946a) and (1952a).]

1940. Review of Louis H[erbert] Gray (1875–1955), *Foundations of Language* (New York: Macmillan, 1939). *Language* 16.3: 216–231. (Repr., with the title "Gray's *Foundations of Language*", in Harris 1970a: 695–705.)

1941a. "Linguistic Structure of Hebrew". *Journal of the American Oriental Society* 61: 143–167. [Also published as *Publications of the American Oriental Society*; Offprint series, No. 14.]

1941b. Review of N[ikolaj] S[ergeevič] Trubetzkoy (1890–1938), *Grundzüge der Phonologie* (Prague: Cercle Linguistique de Prague, 1939). *Language* 17: 345–349. (Repr. in Harris 1970a: 706–711, and in *Phonological Theory: Evolution and current practice* ed. by Valerie Becker Makkai, 301–304. New York: Holt, Rinehart & Winston, 1972; repr., Lake Bluff, Ill.: Jupiter Press, 1978.)

1941–1946. "Cherokee Materials". Manuscript 30(I2.4). [Typed D. and A.D. 620L., 575 slips, 10 discs.] Philadelphia: American Philosophical Society Library.

1942a. "Morpheme Alternants in Linguistic Analysis". *Language* 18.3: 169–180. (Repr. in *Readings in Linguistics* [I]: *The development of descriptive linguistics in America since 1925* [*in later editions: 1925–56*] ed. by Martin Joos [Washington, D.C.: American Council of Learned Societies, 1957; 4th ed., Chicago & London: University of Chicago Press, 1966], pp. 109–115 [with a postscript by Joos, p. 115] and subsequently in Harris 1970a: 78–90, and 1981: 23–35.)

1942b. "Phonologies of African Languages: The phonemes of Moroccan Arabic". *Journal of the American Oriental Society* 62.4: 309–318. (Repr., under the title of "The Phonemes of Moroccan Arabic", in Harris 1970a. 161–176.)

[Read at the Centennial Meeting of the Society, Boston 1942. — Cf. the critique by Jean Cantineau, "Réflexions sur la phonologie de l'arabe marocain", *Hespéris* 37. 193–207 (1951 for 1950). —Harris (2002[1990]) states that "it was possible to describe the entire program from the outset, e.g. in" this paper.]

1942c. Review of *Language, Culture, and Personality: Essays in memory of Edward Sapir* ed. by Leslie Spier, A[lfred] Irving Hallowell & Stanley S[tewart] Newman (Menasha, Wisconsin: Edward Sapir Memorial Fund, 1941). *Language* 18: 238–245.

1942d. (Together with William E[verett] Welmers [1916–1988].) "The Phonemes of Fanti". *Journal of the American Oriental Society* 62: 318–333.[1]

1942e. (Together with Fred Lukoff [1920–2000].) "The Phonemes of Kingwana-Swahili". *Journal of the American Oriental Society* 62: 333–338.

1944a. "Yokuts Structure and [Stanley] Newman's Grammar". *IJAL* 10.4: 196–211. (Repr. in Harris 1970a: 188–208.)

1944b. "Simultaneous Components in Phonology". *Language* 20: 181–205. (Repr. in *Readings in Linguistics* [I]: *The development of descriptive linguistics in America since 1925* [*in later editions: 1925–56*] ed. by Martin Joos [Washington, D.C.: American

1. Harris also served as the editor of *JAOS* from 1941 to 1947.

Council of Learned Societies, 1957; 4th ed., Chicago & London: University of Chicago Press, 1966], pp. 124–138 [with a postscript by Joos, p. 138] and subsequently also in Harris 1970a: 3–31 as well as in *Phonological Theory: Evolution and current practice* ed. by Valerie Becker Makkai, 115–133. New York: Holt, Rinehart & Winston, 1972; repr., Lake Bluff, Ill.: Jupiter Press, 1978.)

1945a. "Navaho Phonology and [Harry] Hoijer's Analysis". *IJAL* 11.4: 239–246. (Repr. in Harris 1970a: 177–187.)

1945b. "Discontinuous Morphemes". *Language* 21.2: 121–127. (Repr. in Harris 1970a: 91–99, and in Harris 1981: 36–44.)

1945c. "American Indian Linguistic Work and the Boas Collection". *Library Bulletin of the American Philosophical Society* 1945: 57–61. Philadelphia.
Review note by Thomas A[lbert] Sebeok (1920–2002) in *IJAL* 13: 126 (1947).

1945d. (Together with Charles F. Voegelin.) *Index to the Franz Boas Collection of Materials for American Linguistics.* (= *Language Monographs*, 22.) Baltimore, Md.: Linguistic Society of America, 43 pp. (Repr., New York: Kraus, 1974.)

1945e. (Together with Charles F. Voegelin.) "Linguistics in Ethnology". *Southwestern Journal of Anthropology* 1: 455–465.

1945f. Review of Murray B[arnson] Emeneau, *Kota Texts*, vol. I (Berkeley: University of California Press, 1944). *Language* 21: 283–289. (Repr., under the title "Emeneau's Kota Texts", in Harris 1970a: 209–216.)
[According to Harris (1990, 2002), in the late 1930s and early 1940s he used the Kota texts that are reviewed here, together with the Hidatsa material of (1939c), for his development of substitution grammar, discourse analysis, and transformational analysis, research which was demonstrated at the Linguistic Institute and subsequently first published in (1946a) and (1952a).]

1946a. "From Morpheme to Utterance". *Language* 22.3: 161–183. (Repr. in Harris 1970a: 100–125, and in Harris 1981: 45–70.)

1946b. (Together with Ernest Bender [b.1919].) "The Phonemes of North Carolina Cherokee". *IJAL* 12: 14–21. (Repr. in *Readings in Linguistics* [I]*: The development of descriptive linguistics in America since 1925* [*in later editions: 1925–56*] ed. by Martin Joos [Washington, D.C.: American Council of Learned Societies, 1957; 4th ed., Chicago & London: University of Chicago Press, 1966], pp. 142–153 [with a postscript by Joos, p. 153].)

1947a. "Developments in American Indian Linguistics". *Library Bulletin of the American Philosophical Society* 1946: 84–97. Philadelphia.
Review note by Thomas A[lbert] Sebeok in *IJAL* 14: 209 (1948).

1947b. "Structural Restatements I: Swadesh's Eskimo; Newman's Yawelmani". *IJAL* 13.1: 47–58. (Repr. in Harris 1970a: 217–234, and in Harris 1981: 71–88.)
["Attempt to restate in summary fashion the grammatical structures of a number of American Indian languages. The languages to be treated are those presented in H[arry] Hoijer and others, *Linguistic Structures of Native America* [New York, 1946]." – On Morris Swadesh's account of Eskimo and Stanley S. Newman's of Yawelmani Yokuts.]

1947c. "Structural Restatements II: Voegelin's Delaware". *IJAL* 13.3: 175–186. (Repr. in Harris 1970a: 235–250, and 1981: 89–104.)
[On Voegelin's grammatical sketch of Delaware.]

1947d. (Together with Charles F. Voegelin.) "The Scope of Linguistics". *American Anthropologist* 49:588–600.
[1, The place of linguistics in cultural anthropology; 2, Trends in linguistics.]

1947e. (Associate ed., with Helen Boas-Yampolsky as main ed.) Franz Boas, *Kwakiutl Grammar, with glossary of the suffixes.* (= *Transactions of the American Philosophical Society*, n.s. 37:199–377.) Philadelphia: American Philosophical Society.
Reviewed by
Morris Swadesh in *Word* 4:58–63 (1948);
C[harles] F[rederick] Voegelin in *Journal of American Folklore* 61:414–415(1948).

1948. "Componential Analysis of a [Modern] Hebrew Paradigm". *Language* 24.1:87–91. (Repr. in *Readings in Linguistics* [I]*: The development of descriptive linguistics in America since 1925* [*in later editions: 1925–56*] ed. by Martin Joos [Washington, D.C.: American Council of Learned Societies, 1957; 4th ed., Chicago & London: University of Chicago Press, 1966], pp. 272–274 [with a postscript by Joos, p. 274] and — with 'Hebrew' in the title dropped — in Harris 1970a:126–130.)

1951a. *Methods in Structural Linguistics.* Chicago: University of Chicago Press, xvi, 384 pp. (Repr., under the title of *Structural Linguistics*, as "Phoenix Books" P 52, 1960; 7th impression, 1966; repr. again in Harris 1984.) [Preface signed "Philadelphia, January 1947".]
Reviewed by
Norman A[nthony] McQuown in *Language* 28.4:495–504 (1952);
Murray Fowler in *Language* 28:504–509 (1952);
C[harles] F[rederick] Voegelin in *Journal of the American Oriental Society* 72:113–114 (1952);
Charles F[rancis] Hockett in *American Speech* 27:117–121 (1952);
Stanley S[tewart] Newman in *American Anthroplogist* 54:404–405 (1952);
Margaret Mead in *IJAL* 18:257–260 (1952);
Fred W[alter] Householder in *IJAL* 18:260–268 (1952);
Fernand Mossé in *Études Germaniques* 7:274 (1952);
Walburga von Raffler[-Engel] in *Paideia* 8:229–230 (1953);
Knud Togeby in *Modern Language Notes* 68:191–194 (1954);
K[enneth] R. Brooks in *Modern Language Review* 48:496 (1953);
Milka Iviç in *JuŬnoslovenski Filolog* 20:474–478 (Belgrade, 1953/54);
Jean Cantineau in *Bulletin de la Société de Linguistique de Paris* 50.2:4–9(1954);
Eugene Dorfman in *Modern Language Journal* 38:159–160 (1954);
Robert Léon Wagner in *Journal de Psychologie* 47:537–539 (1954);
Harry Hoijer in *Romance Philology* 9:32–38 (1955–56);
Paul L[ucian] Garvin in *Romance Philology* 9:38–41 (1955/56).

1951b. (With Charles F. Voegelin.) "Methods for Determining Intelligibility among Dialects of Natural Languages". *Proceedings of the American Philosophical Society* 95: 322–329; 1 fig. Philadelphia.

1951c. Review of David G. Mandelbaum (ed.), *Selected Writings of Edward Sapir in Language, Culture, and Personality* (Berkeley & Los Angeles: University of California Press, 1949). *Language* 27.3:288–333. (Repr. in Harris 1970a:712–764, and in *Edward Sapir:*

Appraisals of his life and work ed. by Konrad Koerner, 69–114. Amsterdam & Philadelphia: John Benjamins, 1984.)

[This insightful review article is also to be reprinted in *Edward Sapir: Critical Assessments* ed. by E. F. K. Koerner, vol. I (London & New York: Routledge, 2003).]

1951d. "Ha-Safah ha-Ivrit l'or ha-balshanut ha-chadashah [The Hebrew language in the light of modern linguistics]". *Lěšonénu: A journal for the study of the Hebrew language and cognate studies* 17: 128–132 (1950/1951). Jerusalem.

1952a. "Culture and Style in Extended Discourse". *Selected Papers from the 29th International Congress of Americanists (New York, 1949)*, vol. III: *Indian Tribes of Aboriginal America* ed. by Sol Tax & Melville J[oyce] Herskovits, 210–215. New York: Cooper Square Publishers. (Entire volume repr., New York: Cooper Press, 1967; paper repr. in Harris 1970: 373–389.)

[Proposes a method for analyzing extended discourse, with sample analyses from Hidatsa, a Siouan language spoken in North Dakota.]

1952b. "Discourse Analysis". *Language* 28.1: 1–30. (Repr. in *The Structure of Language: Readings in the philosophy of language* ed. by Jerry A[lan] Fodor & Jerrold J[acob] Katz, 355–383. Englewood Cliffs, N.J.: Prentice-Hall, 1964; and also in Harris 1970a: 313–348 and Harris 1981: 107–142.)

[Presents a method for the analysis of connected speech or writing.]

1952c. "Discourse Analysis: A sample text". *Language* 28.4: 474–494. (Repr. in Harris 1970a: 349–379.)

1952d. (Together with Charles F. Voegelin.) "Training in Anthropological Linguistics". *American Anthropologist* 54: 322–327.

1953. (Together with C. F. Voegelin.) "Eliciting in Linguistics". *Southwestern Journal of Anthropology* 9.1: 59–75. (Repr. in Harris 1970a: 769–774.)

[1, Practices with respect to eliciting; 2, Imitation and repetition; 3, Eliciting with pictures; 4, Translation eliciting; 5, Text eliciting, and 6, The validity of eliciting.]

1954a. "Transfer Grammar". *IJAL* 20.4: 259–270. (Repr. in Harris 1970a: 139–157.)

[1, "Defining difference between languages"; 2, "Structural transfer"; 3, "Phonetic and phonemic similarity"; 4, "Morphemes and morphophonemes"; 5, "Morphological translatability".]

1954b. "Distributional Structure". *Word* 10:2/3: 146–162. (Also in *Linguistics Today: Published on the occasion of the Columbia University Bicentennial* ed. by André Martinet & Uriel Weinreich, 26–42. New York: Linguistic Circle of New York, 1954. Repr. in *The Structure of Language: Readings in the philosophy of language* ed. by Jerry A[lan] Fodor & Jerrold J[acob] Katz, 33–49. Englewood Cliffs, N.J.: Prentice-Hall, 1964, and also in Harris 1970a: 775–794 and 1981: 3–22.)

1955a. "From Phoneme to Morpheme". *Language* 31.2: 190–222; 7 tables. (Repr. in Harris 1970a: 32–67.)

[Presents a constructional procedure segmenting an utterance in a way which correlates well with word and morpheme boundaries.]

1955b. "American Indian Work and the Boas Collection". *Library Bulletin of the American Philosophical Society* 1955: 57–61. Philadelphia.

1956a. (Editor), *A Bushman Dictionary* by Dorothea F[rances] Bleek [d. 1948]. (= *American Oriental Series*, 41.) New Haven, Conn.: American Oriental Society, xii, 773 pp.
Reviewed by
C[lement] M[artyn] Doke in *African Studies* 16: 124–125 (1957);
E.O.J. Westphal in *Africa* 27: 203–204 (1957);
A. J. C[oetzee] in *Tydskrif vir Volkskunde en Volkstaal* 14.1: 29–30(Johannesburg, 1957);
Joseph H[arold] Greenberg in *Language* 33: 495–497 (1957);
Henri Peter Blok in *Neophilologus* 41: 232–234 (1957);
Louis Deroy in *Revue des Langues Vivantes* 23: 174–175 (1957);
Otto Köhler in *Afrika und Übersee* 43: 133–138 (1959).

1956b. "Introduction to Transformations". (= *Transformations and Discourse Analysis Papers*, No. 2.) Philadelphia: University of Pennsylvania. (Repr. in Harris 1970a: 383–389.)

1957a. "Co-Occurrence and Transformation in Linguistic Structure". *Language* 33.3: 283–340. (Repr. in *The Structure of Language: Readings in the philosophy of language* ed. by Jerry A[lan] Fodor & Jerrold J[acob] Katz, 155–210. Englewood Cliffs, N.J.: Prentice-Hall, 1964, and also in Harris 1970: 390–457, 1972: 78–104 [in parts]. Anthologized in *Syntactic Theory 1: Structuralist. Selected readings* ed. by Fred W. Householder, 151–185. Harmondsworth, Middlesex & Baltimore, Md.: Penguin Books, 1972, and also repr. in Harris 1981: 143–210.)
[Revised and enlarged version of Presidential Address, Linguistic Society of America, December 1955. — Defines a formal relation among sentences, by virtue of which one sentence structure may be called a transform of another sentence structure.]

1957b. "Canonical Form of a Text". (= *Transformations and Discourse Analysis Papers*, No. 3b.) Philadelphia: University of Pennsylvania.
[This and two other previously unpublished papers — items 4a and 3c in the same series — were combined to form entry 1963a (below).]

1959a. "The Transformational Model of Language Structure". *Anthropological Linguistics* 1.1: 27–29.

1959b. "Computable Syntactic Analysis". (= *Transformations and Discourse Analysis Papers*, No. 15.) Philadelphia: University of Pennsylvania. (Revised version published as item 1962a; excerpted, with the added subtitle "The 1959 computer sentence-analyzer", in Harris 1970a: 253–277.)

1959c. *Linguistic Transformations for Information Retrieval.* (= *Interscience Tracts in Pure and Applied Mathematics*, 1958:2.) Washington, D.C.: National Academy of Sciences — National Research Council. (Repr. in Harris 1970a: 458–471.)
[From the 1958 Proceedings of the International Conference on Scientific Information.]

1960a. *Structural Linguistics.* (= *Phoenix Books*, P 52.) Chicago: University of Chicago Press, xvi, 384 pp. (7th impression, 1966; repr., 1984.) [Reprint of item 1951, with a supplementary preface (vi-vii).]
Reviewed by
Simeon Potter in *Modern Language Review* 57: 139 (1962).

1960b. "English Transformation List". (= *Transformations and Discourse Analysis Papers*, No. 30.) Philadelphia: University of Pennsylvania.

1961. "Strings and Transformations in Language Description". Published as No. 1 of *Papers in Formal Linguistics* ed. by Henry Hiż. Department of Linguistics, University of Pennsylvania. (Published, under the title "Introduction to String Analysis", in Harris 1970a: 278–285.)

1962a. *String Analysis of Sentence Structure*. (= *Papers on Formal Linguistics*, 1.) The Hague: Mouton, 70 pp. (2nd ed., 1964; repr., 1965.)
[Revised version of item 1959b.]
Reviewed by
Robert E[dmondson] Longacre in *Language* 39: 473–478 (1963);
László Antal in *Linguistics* No. 1: 97–104 (1963);
Murray Fowler in *Word* 19: 245–247 (1963);
Klaus Baumgärtner in *Germanistik* 4: 194 (1963);
Robert B[enjamin] Lees in *IJAL* 30: 415–420 (1964);
Karel Pala in *Sborník Prací Filosofické Fakulty Brnûnskě Univerzity* 13 (A 12).238–241 (Brno, 1964);
G. G. Pocepkov in *Voprosy Jazykoznania* 13.1: 123–128 (1965);
Karel Pala in *Slovo a Slovesnost* 26: 78–80 (1965);
Kazimierz Polański in *Biuletyn Fonegraficzne* 8: 139–143 (Pozna , 1967).

1962b. "Sovmestnaja vstrecaemost' i transformacija v jazykovoj strukture". *Novoe v lingvistike* ed. by V[ladimir] A[ndreevič] Zvegincev, vol. II: *Transformacionnaja grammatika*, 528–636. Moscow: Izd. Innostr. Literatury. [Transl. by T(atjana) N. Molosaja of item 1957a, with an introd. by S(ebastian) K(onstantinovič) Šaumjan.]

1962c. "A Language for International Cooperation". *Preventing World War III: Some proposals* ed. by Quincy Wright, William M. Evan & Morton Deutsch, 299–309. New York: Simon & Schuster. (Repr. in Harris 1970a. 795–805.)

1963a. *Discourse Analysis Reprints*. (= *Papers on Formal Linguistics*, 2.) The Hague: Mouton, 73 pp. [See comment on entry 1957b (above).]
Reviewed by
Klaus Baumgärtner in *Germanistik* 5: 412 (1964);
Manfred Bierwisch in *Linguistics* No. 13: 61–73 (1965);
Fred[erick] C[hen] C[hung] Peng in *Lingua* 1 6: 325–330 (1966);
György Hell in *Acta Linguistica Academiae Scientarum Hungaricae* 18: 233–235(1968);
Tae-Yong Pak in *Language* 46: 754–764 (1970).

1963b. "Immediate-Constituent Formulation of English Syntax". (= *Transformations and Discourse Analysis Papers*, No. 45.) Philadelphia: University of Pennsylvania. (Repr. in Harris 1970a: 131–138.)

1964a. "Transformations in Linguistic Structure". *Proceedings of the American Philosophical Society* 108.5: 418–422. (Repr. in Harris 1970a: 472–481.) [Read on 25 April 1964.]

1964b. "The Elementary Transformations". (= *Transformations and Discourse Analysis Papers*, No. 54.) Philadelphia: University of Pennsylvania. (Excerpted in Harris 1970a: 482–532, 1972: 57–75, and, in abbreviated form, in Harris 1981: 211–235.)

1965. "Transformational Theory". *Language* 41.3:363–401. (Repr. in Harris 1970a:533–577, 1972:108–154, and 1981:236–280.)

1966a. "Algebraic Operations in Linguistic Structure". Paper read at the International Congress of Mathematicians, Moscow 1966. (Published in Harris 1970a:603–611.)

1966b. "A Cyclic-Cancellation Automation for Sentence Well-Formedness". *International Computation Centre Bulletin* 5:69–94. (Also distributed as *Transformations and Discourse Analysis Papers*, No. 51. Repr. in Harris 1970a:286–309.)

1967a. "Decomposition Lattices". (= *Transformations and Discourse Analysis Papers* No. 70.) Philadelphia: University of Pennsylvania. (Repr. in Harris 1970a. 578–602, and excerpted in Harris 1981:281–290.)

1967b. "Morpheme Boundaries within Words: Report on a computer test". (= *Transformations and Discourse Analysis Papers*, No. 73.) Philadelphia: University of Pennsylvania. (Repr. in Harris 1970a:68–77.)

1968a. *Mathematical Structures of Language*. (= *Interscience Tracts in Pure and Applied Mathematics*, 21.) New York: Interscience Publishers John Wiley & Sons, ix, 230 pp. [Index of terms compiled by Maurice Gross.]

Reviewed by

Wojciech Skalmowski in *ITL: Tijdschrift van het Instituut voor Toegepaste Linguïstiek* 4:56–61 (Leuven, 1969);

Maurice Gross in *Semiotica* 2:380–390 (1970), repr. in item 1972:314–324 (with an introd. in German by Senta Plötz [p. 313] and an English abstract by the author [p. 314]);

Maurice Gross & Marcel-Paul Schützenberger in *The American Scientist* 58(1970); repr. in Harris 1972:308–312 (with summaries in German and English by Senta Plötz [p. 307]);

Petr Pitha in *Slovo a Slovesnost* 32:59–65 (1971);

Lucia Vaina-Puşcă in *Revue Roumaine de Linguistique* 16:369–371 (1971).

1968b. "Edward Sapir: Contributions to linguistics". *International Encyclopedia of the Social Sciences* ed. by David L. Sills, vol. XIV, pp. 13–14. New York: Macmillan. (Repr., in a somewhat longer, probably the original, form in Harris 1970a:765–768.)

1968c. "Du morphème à l'expression". *Langages* No. 9:23–50. [Transl. of item 1946b.]

1969a. *The Two Systems of Grammar: Report and paraphrase*. (= *Transformations and Discourse Analysis Papers*, No. 79.) Philadelphia: University of Pennsylvania. (Repr. in Harris 1970a:612–692, in Harris 1972:158–240 (revised), and in Harris 1981:293–351 (shortened).)

1969b. "Analyse du discours". *Langages* No. 13:8–45. [French transl. of item 1952b.]

1969c. "Mathematical Linguistics". *The Mathematical Sciences* ed. by the Committee on Support of Research in the Mathematical Sciences (COSRIMS), with the collaboration of George A. W. Boehm, 190–196. Cambridge, Mass.: MIT Press.

1970a. *Papers in Structural and Transformational Linguistics*. [Ed. by Henry Hiż.] Dordrecht/Holland: D. Reidel., x, 850 pp.

[Collection of 37 papers originally published between 1940–1969. These are organized under the following headings: 1, "Structural Linguistics, 1: Methods"; 2, "Structural Linguistics, 2: Linguistic structures"; 3, "String Analysis and Computa-

tion"; 4, "Discourse Analysis"; 5, "Transformations", and 6, "About Linguistics". "Preface" (v–vii).]
Reviewed by
Ferenc Kiefer in *Statistical Methods in Linguistics* 7: 60–62 (Stockholm, 1971);
Michael B[enedict] Kac in *Language* 49: 466–473 (1973).

1970b. "La structure distributionnelle". *Analyse distributionnelle et structurale* ed. by Jean Dubois & Françoise Dubois-Charlier (= *Langages*, No. 20), 14–34. Paris: Didier / Larousse. [Transl. of item 1954b.]

1970c. "New Views of Language". Manuscript. (Published in Harris 1972: 242–248, with an introd. in German by the ed. [241–242].)

1971. *Structures mathématiques du langage*. Transl. into French by Catherine Fuchs. (= *Monographies de Linguistique mathématique*, 3.) Paris: Dunod, 248 pp. [Transl. of item 1968a.]
Reviewed by
Yves Gentilhomme in *Bulletin de la Société de Linguistique de Paris* 69.2: 37–53 (1974).

1972. *Transformationelle Analyse: Die Transformationstheorie von Zellig Harris und ihre Entwicklung / Transformational Analysis: The transformational theory of Zellig Harris and its development*. Ed. by Senta Plötz. (= *Linguistische Forschungen*, 8.) Frankfurt/Main: Athenäum-Verlag, viii, 511 pp.
[Reprint of items 1964b (57–75), 1957 (78–104), 1965 (108–154), 1969a (158–240) —revised by the author in 1972, and 1970c (242–248), each introduced, in German, by the ed. (55–57, 76–78, 105–108, 155–157, and 241–242, respectively.]

1973a. "Les deux systèmes de grammaire: Prédicat et paraphrase". *Langages* No. 29: 55–81. [Partial transl., by Danielle Leeman, of item 1969a.]

1973b. Review of *A Leonard Bloomfield Anthology* ed. by Charles F. Hockett (Bloomington & London: Indiana University Press, 1970). *IJAL* 39.4: 252–255.

1976a. "A Theory of Language Structure". *American Philosophical Quarterly* 13: 237–255. (Repr. in Harris 1981: 352–376.)
[Theory of the structure and information of sentences.]

1976b. "On a Theory of Language". *Journal of Philosophy* 73: 253–276. (Excerpted in Harris 1981: 377–391.)

1976c. *Notes du cours de syntaxe*. Transl. and presented by Maurice Gross. Paris: Éditions du Seuil, 236 p. [Transl. of lectures on English syntax given at the Département de Linguistique, University de Paris–Vincennes, 1973–1974.]
Reviewed by
G. L[urquin] in *Le Langage et l'Homme* 31: 114–115 (1976);
Claude Hagège in *Bulletin de la Société de Linguistique de Paris* 72.2: 35–37 (1974);
Riccardo Ambrosini in *Studi e Saggi Linguistici* 17: 309–340 (1977).

1976d. "Morphemalternanten in der linguistischen Analyse". *Beschreibungsmethoden des amerikanischen Strukturalismus* ed. by Elisabeth Bense, Peter Eisenberg & Hartmut Haberland, 129–143. München: Max Hueber. [Transl. by Elisabeth Bense of item 1942a.]

1976e. "Vom Morphem zur Äußerung". *Ibid.*, 181–210. [Transl., by Dietmar Rösler, of item 1946b.]

1976f. "Textanalyse". *Ibid.*, 261–298. [Transl., by Peter Eisenberg, of item 1952b.]

1978a. "Grammar on Mathematical Principles". *Journal of Linguistics* 14:1–20. (Repr. in Harris 1981:392–411.)
["Given as a lecture in Somerville College, Oxford, 16 March 1977".]
1978b. "Operator-Grammar of English". *Lingvisticae Investigationes* 2:55–92. (Excerpted in Harris 1981:412–435.)
1978c. "The Interrogative in a Syntactic Framework". *Questions* ed. by Henry Hiż (= *Synthese Language Library*, 1), 1–35. Dordrecht/Holland: D. Reidel.
1979a. "Założenia metodologiczne językoznawstwa strukturalnego [The methodological basis of structural linguistics]". *Językoznawstwo strukturalne: Wybór tekstów* ed. by Halina Kurkowska & Adam Weinsberg, 158–174. Warsaw: Państwowe Wydawnictwo Naukowe, 274 pp. [Polish transl., by the first editor, of Harris (1951a:4–24), "Methodological Preliminaries".]
1979b. "Mathematical Analysis of Language". Paper delivered to the 6th International Congress on Logic, Methodology, and the Philosophy of Science, held in Hanover, Germany, August 1979. Unpublished.
1981. *Papers on Syntax*. Ed. by Henry Hiż. (= *Synthese Language Library*, 14.) Dordrecht/Holland: D. Reidel, vii, 479 pp.
[Collection of 16 previously published papers, organized under 3 sections: I, "Structural Analysis"; II, "Transformational Analysis", and III, "Operator Grammar". Index (437–479).]
1982a. *A Grammar of English on Mathematical Principles*. New York: John Wiley & Sons, xvi, 429 pp.
Reviewed by
William Frawley in *Language* 60.1:150–152 (1984);
Frank Heny in *Journal of Linguistics* 20.1:181–188 (1984);
Bruce E. Nevin in *Computational Linguistics* 10:3/4:203–211 (1984);
Eric S. Wheeler in *Computers in the Humanities* 17.3:88–92 (1984).
1982b. "Discourse and Sublanguage". *Sublanguage: Studies of language in restricted semantic domains* ed. by Richard Kittredge & John Lehrberger, 231–236. Berlin: Walter de Gruyter.
1985. "On Grammars of Science". *Linguistics and Philosophy: Essays in honor of Rulon S. Wells* ed. by Adam Makkai & Alan K. Melby (= *Current Issues in Linguistic Theory*, 42), 139–148. Amsterdam & Philadelphia: John Benjamins.
1987. "The Structure of Science Information". Paper submitted to the journal *Science*, but rejected by the editor, allegedly because it contained no reference to Chomsky. Unpublished.
1988a. *Language and Information*. (= *Bampton Lectures in America*, 28.) New York: Columbia University Press, ix, 120 pp.
[Revised version of lectures given at Columbia University, New York City, in Oct. 1986. — 1, "A Formal Theory of Syntax"; 2, "Scientific Sub-Languages"; 3, "Information", and 4, "The Nature of Language".]
Reviewed by
P[eter] H[ugoe] Matthews in *Times Literary Supplement* (London, 23–29 Dec. 1988), with the title "Saying Something Simple".

1988b. (Together with Paul Mattick, Jr.) "Scientific Sublanguages and the Prospects for a Global Language of Science". *Annals of the American Association of Philosophy and Social Sciences* No. 495: 73–83.

1989. (Together with Michael Gottfried, Thomas Ryckman, Paul Mattick, Jr., Anne Daladier, Tzvee N. Harris & Suzanna Harris.) *The Form of Information in Science: Analysis of an immunology sublanguage*. Preface by Hilary Putnam. (= *Boston Studies in the Philosophy of Science*, 104.) Dordrecht/Holland & Boston: Kluwer Academic Publishers, xvii, 590 pp.

1990. "La genèse de l'analyse des transformations et de la métalangue". *Langages* No. 99 (Sept. 1990), 9–19. [Transl. of item 2002 by Anne Daladier.]

1991. *A Theory of Language and Information: A mathematical approach*. Oxford & New York: Clarendon Press, xii, 428 pp.; illustr.

Reviewed by

Jorge Baptista in *Revista da Faculdade de Letras* 15 (5ª, Série): 203–205. Lisboa: Faculdade de Letras da Universidade de Lisboa (FLUL);

D. Terence Langendoen in *Language* 70.3: 585–588 (1994).

1997. *The Transformation of Capitalist Society*. Foreword by Wolf V. Heydebrand [xi-xiii]. Baltimore, Md.: Rowman & Littlefield, xvi, 244 pp. (and an unnumbered page "About the Author").

["On Behalf of the Author" (signed by Murray Eden, William M. Evan, Seymour Melman) concludes with the sentence "Several of his old friends collaborated in preparing the manuscript for publication." "Preface" by Zellig S. Harris (xv-xvi). Contents: Chap. 1, "Overview: The possibilities of change" (1–7), with Appendix "Criticizing capitalist society" (9–10); Chap. 2, "Basic terms in describing society" (11–22); Chap. 3, "Capitalist decisions on production" (23–42); Chap. 4, "Considerations in analyzing social change" (43–55); Chap. 5, "Potentially post-capitalist developments" (57–86); Chap. 6, "How capitalism began" (87–112); Chap. 7, "Self-governed production" (113–182); Chap. 8, "In the aftermath of Soviet communism" (183–208), and Chap. 9, "Intervening in the historical process" (209–233). Index (235–244).]

Reviewed by

Peter Franz in *The European Legacy* 3: 112–113 (1998).

2002. "The Background of Transformational and Metalanguage Analysis". *The Legacy of Zellig Harris: Language and information into the 21st century*, Volume I: *Philosophy of science, syntax, and semantics* ed. by Bruce E. Nevin (= *Current Issues in Linguistic Theory*, 228), 1–14. Amsterdam & Philadelphia: John Benjamins.

[Publication of original English text with portions not included in item 1990. — Proposes a method for analyzing extended discourse, with sample analyses from Hidatsa, a Siouan language spoken in North Dakota.]

Name index

Subject index

CURRENT ISSUES IN LINGUISTIC THEORY

E. F. K. Koerner, Editor
Institut für Sprachwissenschaft, Universität zu Köln
D-50923 KÖLN, Germany
efk.koerner@uni-koeln.de

The *Current Issues in Linguistic Theory* (CILT) series is a theory-oriented series which welcomes contributions from scholars who have significant proposals to make towards the advancement of our understanding of language, its structure, functioning and development. CILT has been established in order to provide a forum for the presentation and discussion of linguistic opinions of scholars who do not necessarily accept the prevailing mode of thought in linguistic science. It offers an alternative outlet for meaningful contributions to the current linguistic debate, and furnishes the diversity of opinion which a healthy discipline must have. In this series the following volumes have been published thus far or are scheduled for publication:

1. KOERNER, Konrad (ed.): *The Transformational-Generative Paradigm and Modern Linguistic Theory.* 1975.
2. WEIDERT, Alfons: *Componential Analysis of Lushai Phonology.* 1975.
3. MAHER, J. Peter: *Papers on Language Theory and History I: Creation and Tradition in Language. Foreword by Raimo Anttila.* 1979.
4. HOPPER, Paul J. (ed.): *Studies in Descriptive and Historical Linguistics. Festschrift for Winfred P. Lehmann.* 1977.
5. ITKONEN, Esa: *Grammatical Theory and Metascience: A critical investigation into the methodological and philosophical foundations of 'autonomous' linguistics.* 1978.
6. ANTTILA, Raimo: *Historical and Comparative Linguistics.* 1989.
7. MEISEL, Jürgen M. & Martin D. PAM (eds): *Linear Order and Generative Theory.* 1979.
8. WILBUR, Terence H.: *Prolegomena to a Grammar of Basque.* 1979.
9. HOLLIEN, Harry & Patricia (eds): *Current Issues in the Phonetic Sciences. Proceedings of the IPS-77 Congress, Miami Beach, Florida, 17-19 December 1977.* 1979.
10. PRIDEAUX, Gary D. (ed.): *Perspectives in Experimental Linguistics. Papers from the University of Alberta Conference on Experimental Linguistics, Edmonton, 13-14 Oct. 1978.* 1979.
11. BROGYANYI, Bela (ed.): *Studies in Diachronic, Synchronic, and Typological Linguistics: Festschrift for Oswald Szemérenyi on the Occasion of his 65th Birthday.* 1979.
12. FISIAK, Jacek (ed.): *Theoretical Issues in Contrastive Linguistics.* 1981. Out of print
13. MAHER, J. Peter, Allan R. BOMHARD & Konrad KOERNER (eds): *Papers from the Third International Conference on Historical Linguistics, Hamburg, August 22-26 1977.* 1982.
14. TRAUGOTT, Elizabeth C., Rebecca LaBRUM & Susan SHEPHERD (eds): *Papers from the Fourth International Conference on Historical Linguistics, Stanford, March 26-30 1979.* 1980.
15. ANDERSON, John (ed.): *Language Form and Linguistic Variation. Papers dedicated to Angus McIntosh.* 1982.
16. ARBEITMAN, Yoël L. & Allan R. BOMHARD (eds): *Bono Homini Donum: Essays in Historical Linguistics, in Memory of J.Alexander Kerns.* 1981.
17. LIEB, Hans-Heinrich: *Integrational Linguistics. 6 volumes. Vol. II-VI n.y.p.* 1984/93.
18. IZZO, Herbert J. (ed.): *Italic and Romance. Linguistic Studies in Honor of Ernst Pulgram.* 1980.
19. RAMAT, Paolo et al. (eds): *Linguistic Reconstruction and Indo-European Syntax. Proceedings of the Colloquium of the 'Indogermanischhe Gesellschaft'. University of Pavia, 6-7 September 1979.* 1980.
20. NORRICK, Neal R.: *Semiotic Principles in Semantic Theory.* 1981.
21. AHLQVIST, Anders (ed.): *Papers from the Fifth International Conference on Historical Linguistics, Galway, April 6-10 1981.* 1982.

22. UNTERMANN, Jürgen & Bela BROGYANYI (eds): *Das Germanische und die Rekonstruktion der Indogermanischen Grundsprache. Akten des Freiburger Kolloquiums der Indogermanischen Gesellschaft, Freiburg, 26-27 Februar 1981.* 1984.
23. DANIELSEN, Niels: *Papers in Theoretical Linguistics. Edited by Per Baerentzen.* 1992.
24. LEHMANN, Winfred P. & Yakov MALKIEL (eds): *Perspectives on Historical Linguistics. Papers from a conference held at the meeting of the Language Theory Division, Modern Language Assn., San Francisco, 27-30 December 1979.* 1982.
25. ANDERSEN, Paul Kent: *Word Order Typology and Comparative Constructions.* 1983.
26. BALDI, Philip (ed.): *Papers from the XIIth Linguistic Symposium on Romance Languages, Univ. Park, April 1-3, 1982.* 1984.
27. BOMHARD, Alan R.: *Toward Proto-Nostratic. A New Approach to the Comparison of Proto-Indo-European and Proto-Afroasiatic. Foreword by Paul J. Hopper.* 1984.
28. BYNON, James (ed.): *Current Progress in Afro-Asiatic Linguistics: Papers of the Third International Hamito-Semitic Congress, London, 1978.* 1984.
29. PAPROTTÉ, Wolf & René DIRVEN (eds): *The Ubiquity of Metaphor: Metaphor in language and thought.* 1985 (publ. 1986).
30. HALL, Robert A. Jr.: *Proto-Romance Morphology. = Comparative Romance Grammar, vol. III.* 1984.
31. GUILLAUME, Gustave: *Foundations for a Science of Language.*
32. COPELAND, James E. (ed.): *New Directions in Linguistics and Semiotics.* Co-edition with Rice University Press who hold exclusive rights for US and Canada. 1984.
33. VERSTEEGH, Kees: *Pidginization and Creolization. The Case of Arabic.* 1984.
34. FISIAK, Jacek (ed.): *Papers from the VIth International Conference on Historical Linguistics, Poznan, 22-26 August. 1983.* 1985.
35. COLLINGE, N.E.: *The Laws of Indo-European.* 1985.
36. KING, Larry D. & Catherine A. MALEY (eds): *Selected papers from the XIIIth Linguistic Symposium on Romance Languages, Chapel Hill, N.C., 24-26 March 1983.* 1985.
37. GRIFFEN, T.D.: *Aspects of Dynamic Phonology.* 1985.
38. BROGYANYI, Bela & Thomas KRÖMMELBEIN (eds): *Germanic Dialects:Linguistic and Philological Investigations.* 1986.
39. BENSON, James D., Michael J. CUMMINGS, & William S. GREAVES (eds): *Linguistics in a Systemic Perspective.* 1988.
40. FRIES, Peter Howard (ed.) in collaboration with Nancy M. Fries: *Toward an Understanding of Language: Charles C. Fries in Perspective.* 1985.
41. EATON, Roger, et al. (eds): *Papers from the 4th International Conference on English Historical Linguistics, April 10-13, 1985.* 1985.
42. MAKKAI, Adam & Alan K. MELBY (eds): *Linguistics and Philosophy. Festschrift for Rulon S. Wells.* 1985 (publ. 1986).
43. AKAMATSU, Tsutomu: *The Theory of Neutralization and the Archiphoneme in Functional Phonology.* 1988.
44. JUNGRAITHMAYR, Herrmann & Walter W. MUELLER (eds): *Proceedings of the Fourth International Hamito-Semitic Congress.* 1987.
45. KOOPMAN, W.F., F.C. Van der LEEK , O. FISCHER & R. EATON (eds): *Explanation and Linguistic Change.* 1986
46. PRIDEAUX, Gary D. & William J. BAKER: *Strategies and Structures: The processing of relative clauses.* 1987.
47. LEHMANN, Winfred P. (ed.): *Language Typology 1985. Papers from the Linguistic Typology Symposium, Moscow, 9-13 Dec. 1985.* 1986.
48. RAMAT, Anna G., Onofrio CARRUBA and Giuliano BERNINI (eds): *Papers from the 7th International Conference on Historical Linguistics.* 1987.
49. WAUGH, Linda R. and Stephen RUDY (eds): *New Vistas in Grammar: Invariance and*

Variation. Proceedings of the Second International Roman Jakobson Conference, New York University, Nov.5-8, 1985. 1991.

50. RUDZKA-OSTYN, Brygida (ed.): *Topics in Cognitive Linguistics.* 1988.
51. CHATTERJEE, Ranjit: *Aspect and Meaning in Slavic and Indic. With a foreword by Paul Friedrich.* 1989.
52. FASOLD, Ralph W. & Deborah SCHIFFRIN (eds): *Language Change and Variation.* 1989.
53. SANKOFF, David: *Diversity and Diachrony.* 1986.
54. WEIDERT, Alfons: *Tibeto-Burman Tonology. A comparative analysis.* 1987
55. HALL, Robert A. Jr.: *Linguistics and Pseudo-Linguistics.* 1987.
56. HOCKETT, Charles F.: *Refurbishing our Foundations. Elementary linguistics from an advanced point of view.* 1987.
57. BUBENIK, Vít: *Hellenistic and Roman Greece as a Sociolinguistic Area.* 1989.
58. ARBEITMAN, Yoël. L. (ed.): *Fucus: A Semitic/Afrasian Gathering in Remembrance of Albert Ehrman.* 1988.
59. VAN VOORST, Jan: *Event Structure.* 1988.
60. KIRSCHNER, Carl & Janet DECESARIS (eds): *Studies in Romance Linguistics. Selected Proceedings from the XVII Linguistic Symposium on Romance Languages.* 1989.
61. CORRIGAN, Roberta L., Fred ECKMAN & Michael NOONAN (eds): *Linguistic Categorization. Proceedings of an International Symposium in Milwaukee, Wisconsin, April 10-11, 1987.* 1989.
62. FRAJZYNGIER, Zygmunt (ed.): *Current Progress in Chadic Linguistics. Proceedings of the International Symposium on Chadic Linguistics, Boulder, Colorado, 1-2 May 1987.* 1989.
63. EID, Mushira (ed.): *Perspectives on Arabic Linguistics I. Papers from the First Annual Symposium on Arabic Linguistics.* 1990.
64. BROGYANYI, Bela (ed.): *Prehistory, History and Historiography of Language, Speech, and Linguistic Theory. Papers in honor of Oswald Szemérenyi I.* 1992.
65. ADAMSON, Sylvia, Vivien A. LAW, Nigel VINCENT and Susan WRIGHT (eds): *Papers from the 5th International Conference on English Historical Linguistics.* 1990.
66. ANDERSEN, Henning and Konrad KOERNER (eds): *Historical Linguistics 1987.Papers from the 8th International Conference on Historical Linguistics,Lille, August 30-Sept., 1987.* 1990.
67. LEHMANN, Winfred P. (ed.): *Language Typology 1987. Systematic Balance in Language. Papers from the Linguistic Typology Symposium, Berkeley, 1-3 Dec 1987.* 1990.
68. BALL, Martin, James FIFE, Erich POPPE &Jenny ROWLAND (eds): *Celtic Linguistics/ Ieithyddiaeth Geltaidd. Readings in the Brythonic Languages. Festschrift for T. Arwyn Watkins.* 1990.
69. WANNER, Dieter and Douglas A. KIBBEE (eds): *New Analyses in Romance Linguistics. Selected papers from the Linguistic Symposium on Romance Languages XVIIII, Urbana-Champaign, April 7-9, 1988.* 1991.
70. JENSEN, John T.: *Morphology. Word structure in generative grammar.* 1990.
71. O'GRADY, William: *Categories and Case. The sentence structure of Korean.* 1991.
72. EID, Mushira and John MCCARTHY (eds): *Perspectives on Arabic Linguistics II. Papers from the Second Annual Symposium on Arabic Linguistics.* 1990.
73. STAMENOV, Maxim (ed.): *Current Advances in Semantic Theory.* 1991.
74. LAEUFER, Christiane and Terrell A. MORGAN (eds): *Theoretical Analyses in Romance Linguistics.* 1991.
75. DROSTE, Flip G. and John E. JOSEPH (eds): *Linguistic Theory and Grammatical Description. Nine Current Approaches.* 1991.
76. WICKENS, Mark A.: *Grammatical Number in English Nouns. An empirical and theoretical account.* 1992.
77. BOLTZ, William G. and Michael C. SHAPIRO (eds): *Studies in the Historical Phonology of Asian Languages.* 1991.

78. KAC, Michael: *Grammars and Grammaticality.* 1992.
79. ANTONSEN, Elmer H. and Hans Henrich HOCK (eds): *STAEF-CRAEFT: Studies in Germanic Linguistics. Select papers from the First and Second Symposium on Germanic Linguistics, University of Chicago, 24 April 1985, and Univ. of Illinois at Urbana-Champaign, 3-4 Oct. 1986.* 1991.
80. COMRIE, Bernard and Mushira EID (eds): *Perspectives on Arabic Linguistics III. Papers from the Third Annual Symposium on Arabic Linguistics.* 1991.
81. LEHMANN, Winfred P. and H.J. HEWITT (eds): *Language Typology 1988. Typological Models in the Service of Reconstruction.* 1991.
82. VAN VALIN, Robert D. (ed.): *Advances in Role and Reference Grammar.* 1992.
83. FIFE, James and Erich POPPE (eds): *Studies in Brythonic Word Order.* 1991.
84. DAVIS, Garry W. and Gregory K. IVERSON (eds): *Explanation in Historical Linguistics.* 1992.
85. BROSELOW, Ellen, Mushira EID and John McCARTHY (eds): *Perspectives on Arabic Linguistics IV. Papers from the Annual Symposium on Arabic Linguistics.* 1992.
86. KESS, Joseph F.: *Psycholinguistics. Psychology, linguistics, and the study of natural language.* 1992.
87. BROGYANYI, Bela and Reiner LIPP (eds): *Historical Philology: Greek, Latin, and Romance. Papers in honor of Oswald Szemerényi II.* 1992.
88. SHIELDS, Kenneth: *A History of Indo-European Verb Morphology.* 1992.
89. BURRIDGE, Kate: *Syntactic Change in Germanic. A study of some aspects of language change in Germanic with particular reference to Middle Dutch.* 1992.
90. KING, Larry D.: *The Semantic Structure of Spanish. Meaning and grammatical form.* 1992.
91. HIRSCHBÜHLER, Paul and Konrad KOERNER (eds): *Romance Languages and Modern Linguistic Theory. Selected papers from the XX Linguistic Symposium on Romance Languages,University of Ottawa, April 10-14, 1990.* 1992.
92. POYATOS, Fernando: *Paralanguage: A linguistic and interdisciplinary approach to interactive speech and sounds.* 1992.
93. LIPPI-GREEN, Rosina (ed.): *Recent Developments in Germanic Linguistics.* 1992.
94. HAGÈGE, Claude: *The Language Builder. An essay on the human signature in linguistic morphogenesis.* 1992.
95. MILLER, D. Gary: *Complex Verb Formation.* 1992.
96. LIEB, Hans-Heinrich (ed.): *Prospects for a New Structuralism.* 1992.
97. BROGYANYI, Bela & Reiner LIPP (eds): *Comparative-Historical Linguistics: Indo-European and Finno-Ugric. Papers in honor of Oswald Szemerényi III.* 1992.
98. EID, Mushira & Gregory K. IVERSON: *Principles and Prediction: The analysis of natural language.* 1993.
99. JENSEN, John T.: *English Phonology.* 1993.
100. MUFWENE, Salikoko S. and Lioba MOSHI (eds): *Topics in African Linguistics. Papers from the XXI Annual Conference on African Linguistics, University of Georgia, April 1990.* 1993.
101. EID, Mushira & Clive HOLES (eds): *Perspectives on Arabic Linguistics V. Papers from the Fifth Annual Symposium on Arabic Linguistics.* 1993.
102. DAVIS, Philip W. (ed.): *Alternative Linguistics. Descriptive and theoretical Modes.* 1995.
103. ASHBY, William J., Marianne MITHUN, Giorgio PERISSINOTTO and Eduardo RAPOSO: *Linguistic Perspectives on Romance Languages. Selected papers from the XXI Linguistic Symposium on Romance Languages, Santa Barbara, February 21-24, 1991.* 1993.
104. KURZOVÁ, Helena: *From Indo-European to Latin. The evolution of a morphosyntactic type.* 1993.
105. HUALDE, José Ignacio and Jon ORTIZ DE URBANA (eds): *Generative Studies in Basque Linguistics.* 1993.
106. AERTSEN, Henk and Robert J. JEFFERS (eds): *Historical Linguistics 1989. Papers from the 9th International Conference on Historical Linguistics, New Brunswick, 14-18 August 1989.* 1993.

107. MARLE, Jaap van (ed.): *Historical Linguistics 1991. Papers from the 10th International Conference on Historical Linguistics, Amsterdam, August 12-16, 1991.* 1993.
108. LIEB, Hans-Heinrich: *Linguistic Variables. Towards a unified theory of linguistic variation.* 1993.
109. PAGLIUCA, William (ed.): *Perspectives on Grammaticalization.* 1994.
110. SIMONE, Raffaele (ed.): *Iconicity in Language.* 1995.
111. TOBIN, Yishai: *Invariance, Markedness and Distinctive Feature Analysis. A contrastive study of sign systems in English and Hebrew.* 1994.
112. CULIOLI, Antoine: *Cognition and Representation in Linguistic Theory. Translated, edited and introduced by Michel Liddle.* 1995.
113. FERNÁNDEZ, Francisco, Miguel FUSTER and Juan Jose CALVO (eds): *English Historical Linguistics 1992. Papers from the 7th International Conference on English Historical Linguistics, Valencia, 22-26 September 1992.*1994.
114. EGLI, U., P. PAUSE, Chr. SCHWARZE, A. von STECHOW, G. WIENOLD (eds): *Lexical Knowledge in the Organisation of Language.* 1995.
115. EID, Mushira, Vincente CANTARINO and Keith WALTERS (eds): *Perspectives on Arabic Linguistics. Vol. VI. Papers from the Sixth Annual Symposium on Arabic Linguistics.* 1994.
116. MILLER, D. Gary: *Ancient Scripts and Phonological Knowledge.* 1994.
117. PHILIPPAKI-WARBURTON, I., K. NICOLAIDIS and M. SIFIANOU (eds): *Themes in Greek Linguistics. Papers from the first International Conference on Greek Linguistics, Reading, September 1993.* 1994.
118. HASAN, Ruqaiya and Peter H. FRIES (eds): *On Subject and Theme. A discourse functional perspective.* 1995.
119. LIPPI-GREEN, Rosina: *Language Ideology and Language Change in Early Modern German. A sociolinguistic study of the consonantal system of Nuremberg.* 1994.
120. STONHAM, John T. : *Combinatorial Morphology.* 1994.
121. HASAN, Ruqaiya, Carmel CLORAN and David BUTT (eds): *Functional Descriptions. Theorie in practice.* 1996.
122. SMITH, John Charles and Martin MAIDEN (eds): *Linguistic Theory and the Romance Languages.* 1995.
123. AMASTAE, Jon, Grant GOODALL, Mario MONTALBETTI and Marianne PHINNEY: *Contemporary Research in Romance Linguistics. Papers from the XXII Linguistic Symposium on Romance Languages, El Paso//Juárez, February 22-24, 1994.* 1995.
124. ANDERSEN, Henning: *Historical Linguistics 1993. Selected papers from the 11th International Conference on Historical Linguistics, Los Angeles, 16-20 August 1993.* 1995.
125. SINGH, Rajendra (ed.): *Towards a Critical Sociolinguistics.* 1996.
126. MATRAS, Yaron (ed.): *Romani in Contact. The history, structure and sociology of a language.* 1995.
127. GUY, Gregory R., Crawford FEAGIN, Deborah SCHIFFRIN and John BAUGH (eds): *Towards a Social Science of Language. Papers in honor of William Labov. Volume 1: Variation and change in language and society.* 1996.
128. GUY, Gregory R., Crawford FEAGIN, Deborah SCHIFFRIN and John BAUGH (eds): *Towards a Social Science of Language. Papers in honor of William Labov. Volume 2: Social interaction and discourse structures.* 1997.
129. LEVIN, Saul: *Semitic and Indo-European: The Principal Etymologies. With observations on Afro-Asiatic.* 1995.
130. EID, Mushira (ed.) *Perspectives on Arabic Linguistics. Vol. VII. Papers from the Seventh Annual Symposium on Arabic Linguistics.* 1995.
131. HUALDE, Jose Ignacio, Joseba A. LAKARRA and R.L. Trask (eds): *Towards a History of the Basque Language.* 1995.
132. HERSCHENSOHN, Julia: *Case Suspension and Binary Complement Structure in French.* 1996.

133. ZAGONA, Karen (ed.): *Grammatical Theory and Romance Languages. Selected papers from the 25th Linguistic Symposium on Romance Languages (LSRL XXV) Seattle, 2-4 March 1995.* 1996.
134. EID, Mushira (ed.): *Perspectives on Arabic Linguistics Vol. VIII. Papers from the Eighth Annual Symposium on Arabic Linguistics.* 1996.
135. BRITTON Derek (ed.): *Papers from the 8th International Conference on English Historical Linguistics.* 1996.
136. MITKOV, Ruslan and Nicolas NICOLOV (eds): *Recent Advances in Natural Language Processing.* 1997.
137. LIPPI-GREEN, Rosina and Joseph C. SALMONS (eds): *Germanic Linguistics. Syntactic and diachronic.* 1996.
138. SACKMANN, Robin (ed.): *Theoretical Linguistics and Grammatical Description.* 1996.
139. BLACK, James R. and Virginia MOTAPANYANE (eds): *Microparametric Syntax and Dialect Variation.* 1996.
140. BLACK, James R. and Virginia MOTAPANYANE (eds): *Clitics, Pronouns and Movement.* 1997.
141. EID, Mushira and Dilworth PARKINSON (eds): *Perspectives on Arabic Linguistics Vol. IX. Papers from the Ninth Annual Symposium on Arabic Linguistics, Georgetown University, Washington D.C., 1995.* 1996.
142. JOSEPH, Brian D. and Joseph C. SALMONS (eds): *Nostratic. Sifting the evidence.* 1998.
143. ATHANASIADOU, Angeliki and René DIRVEN (eds): *On Conditionals Again.* 1997.
144. SINGH, Rajendra (ed): *Trubetzkoy's Orphan. Proceedings of the Montréal Roundtable "Morphophonology: contemporary responses (Montréal, October 1994).* 1996.
145. HEWSON, John and Vit BUBENIK: *Tense and Aspect in Indo-European Languages. Theory, typology, diachrony.* 1997.
146. HINSKENS, Frans, Roeland VAN HOUT and W. Leo WETZELS (eds): *Variation, Change, and Phonological Theory.* 1997.
147. HEWSON, John: *The Cognitive System of the French Verb.* 1997.
148. WOLF, George and Nigel LOVE (eds): *Linguistics Inside Out. Roy Harris and his critics.* 1997.
149. HALL, T. Alan: *The Phonology of Coronals.* 1997.
150. VERSPOOR, Marjolijn, Kee Dong LEE and Eve SWEETSER (eds): *Lexical and Syntactical Constructions and the Construction of Meaning. Proceedings of the Bi-annual ICLA meeting in Albuquerque, July 1995.* 1997.
151. LIEBERT, Wolf-Andreas, Gisela REDEKER and Linda WAUGH (eds): *Discourse and Perspectives in Cognitive Linguistics.* 1997.
152. HIRAGA, Masako, Chris SINHA and Sherman WILCOX (eds): *Cultural, Psychological and Typological Issues in Cognitive Linguistics.* 1999.
153. EID, Mushira and Robert R. RATCLIFFE (eds): *Perspectives on Arabic Linguistics Vol. X. Papers from the Tenth Annual Symposium on Arabic Linguistics, Salt Lake City, 1996.* 1997.
154. SIMON-VANDENBERGEN, Anne-Marie, Kristin DAVIDSE and Dirk NOËL (eds): *Reconnecting Language. Morphology and Syntax in Functional Perspectives.* 1997.
155. FORGET, Danielle, Paul HIRSCHBÜHLER, France MARTINEAU and María-Luisa RIVERO (eds): *Negation and Polarity. Syntax and semantics. Selected papers from the Colloquium Negation: Syntax and Semantics. Ottawa, 11-13 May 1995.* 1997.
156. MATRAS, Yaron, Peter BAKKER and Hristo KYUCHUKOV (eds): *The Typology and Dialectology of Romani.* 1997.
157. LEMA, José and Esthela TREVIÑO (eds): *Theoretical Analyses on Romance Languages. Selected papers from the 26th Linguistic Symposium on Romance Languages (LSRL XXVI), Mexico City, 28-30 March, 1996.* 1998.
158. SÁNCHEZ MACARRO, Antonia and Ronald CARTER (eds): *Linguistic Choice across Genres. Variation in spoken and written English.* 1998.

159. JOSEPH, Brian D., Geoffrey C. HORROCKS and Irene PHILIPPAKI-WARBURTON (eds): *Themes in Greek Linguistics II.* 1998.
160. SCHWEGLER, Armin, Bernard TRANEL and Myriam URIBE-ETXEBARRIA (eds): *Romance Linguistics: Theoretical Perspectives. Selected papers from the 27th Linguistic Symposium on Romance Languages (LSRL XXVII), Irvine, 20-22 February, 1997.* 1998.
161. SMITH, John Charles and Delia BENTLEY (eds): *Historical Linguistics 1995. Volume 1: Romance and general linguistics.* 2000.
162. HOGG, Richard M. and Linda van BERGEN (eds): *Historical Linguistics 1995. Volume 2: Germanic linguistics.Selected papers from the 12th International Conference on Historical Linguistics, Manchester, August 1995.* 1998.
163. LOCKWOOD, David G., Peter H. FRIES and James E. COPELAND (eds): *Functional Approaches to Language, Culture and Cognition.* 2000.
164. SCHMID, Monika, Jennifer R. AUSTIN and Dieter STEIN (eds): *Historical Linguistics 1997. Selected papers from the 13th International Conference on Historical Linguistics, Düsseldorf, 10-17 August 1997.* 1998.
165. BUBENÍK, Vit: *A Historical Syntax of Late Middle Indo-Aryan (Apabhramśa).* 1998.
166. LEMMENS, Maarten: *Lexical Perspectives on Transitivity and Ergativity. Causative constructions in English.* 1998.
167. BENMAMOUN, Elabbas, Mushira EID and Niloofar HAERI (eds): *Perspectives on Arabic Linguistics Vol. XI. Papers from the Eleventh Annual Symposium on Arabic Linguistics, Atlanta, 1997.* 1998.
168. RATCLIFFE, Robert R.: *The "Broken" Plural Problem in Arabic and Comparative Semitic. Allomorphy and analogy in non-concatenative morphology.* 1998.
169. GHADESSY, Mohsen (ed.): *Text and Context in Functional Linguistics.* 1999.
170. LAMB, Sydney M.: *Pathways of the Brain. The neurocognitive basis of language.* 1999.
171. WEIGAND, Edda (ed.): *Contrastive Lexical Semantics.* 1998.
172. DIMITROVA-VULCHANOVA, Mila and Lars HELLAN (eds): *Topics in South Slavic Syntax and Semantics.* 1999.
173. TREVIÑO, Esthela and José LEMA (eds): Semantic Issues in Romance Syntax. 1999.
174. HALL, T. Alan and Ursula KLEINHENZ (eds): *Studies on the Phonological Word.* 1999.
175. GIBBS, Ray W. and Gerard J. STEEN (eds): *Metaphor in Cognitive Linguistics. Selected papers from the 5th International Cognitive Linguistics Conference, Amsterdam, 1997.* 2001.
176. VAN HOEK, Karen, Andrej KIBRIK and Leo NOORDMAN (eds): *Discourse in Cognitive Linguistics. Selected papers from the International Cognitive Linguistics Conference, Amsterdam, July 1997.* 1999.
177. CUYCKENS, Hubert and Britta ZAWADA (eds): *Polysemy in Cognitive Linguistics. Selected papers from the International Cognitive Linguistics Conference, Amsterdam, 1997.* 2001.
178. FOOLEN, Ad and Frederike van der LEEK (eds): *Constructions in Cognitive Linguistics. Selected papers from the Fifth International Cognitive Linguistic Conference, Amsterdam, 1997.* 2000.
179. RINI, Joel: *Exploring the Role of Morphology in the Evolution of Spanish.* 1999.
180. MEREU, Lunella (ed.): *Boundaries of Morphology and Syntax.* 1999.
181. MOHAMMAD, Mohammad A.: *Word Order, Agreement and Pronominalization in Standard and Palestinian Arabic.* 2000.
182. KENESEI, István (ed.): *Theoretical Issues in Eastern European Languages. Selected papers from the Conference on Linguistic Theory in Eastern European Languages (CLITE), Szeged, April 1998.* 1999.
183. CONTINI-MORAVA, Ellen and Yishai TOBIN (eds): *Between Grammar and Lexicon.* 2000.
184. SAGART, Laurent: *The Roots of Old Chinese.* 1999.
185. AUTHIER, J.-Marc, Barbara E. BULLOCK, Lisa A. REED (eds): *Formal Perspectives on Romance Linguistics. Selected papers from the 28th Linguistic Symposium on Romance Languages (LSRL XXVIII), University Park, 16-19 April 1998.* 1999.

186. MIŠESKA TOMIĆ, Olga and Milorad RADOVANOVIĆ (eds): *History and Perspectives of Language Study*. 2000.
187. FRANCO, Jon, Alazne LANDA and Juan MARTÍN (eds): *Grammatical Analyses in Basque and Romance Linguistics*. 1999.
188. VanNESS SIMMONS, Richard: *Chinese Dialect Classification. A comparative approach to Harngjou, Old Jintarn, and Common Northern Wu*. 1999.
189. NICHOLOV, Nicolas and Ruslan MITKOV (eds): *Recent Advances in Natural Language Processing II. Selected papers from RANLP '97*. 2000.
190. BENMAMOUN, Elabbas (ed.): *Perspectives on Arabic Linguistics Vol. XII. Papers from the Twelfth Annual Symposium on Arabic Linguistics*. 1999.
191. SIHLER, Andrew L.: *Language Change. An introduction*. 2000.
192. ALEXANDROVA, Galina M. and Olga ARNAUDOVA (eds.): *The Minimalist Parameter. Selected papers from the Open Linguistics Forum, Ottawa, 21-23 March 1997*. 2001.
193. KLAUSENBURGER, Jurgen: *Grammaticalization. Studies in Latin and Romance morphosyntax*. 2000.
194. COLEMAN, Julie and Christian J. KAY (eds): *Lexicology, Semantics and Lexicography. Selected papers from the Fourth G. L. Brook Symposium, Manchester, August 1998*. 2000.
195. HERRING, Susan C., Pieter van REENEN and Lene SCHØSLER (eds): *Textual Parameters in Older Languages*. 2000.
196. HANNAHS, S. J. and Mike DAVENPORT (eds): *Issues in Phonological Structure. Papers from an International Workshop*. 1999.
197. COOPMANS, Peter, Martin EVERAERT and Jane GRIMSHAW (eds): *Lexical Specification and Insertion*. 2000.
198. NIEMEIER, Susanne and René DIRVEN (eds): *Evidence for Linguistic Relativity*. 2000.
199. PÜTZ, Martin and Marjolijn H. VERSPOOR (eds): *Explorations in Linguistic Relativity*. 2000.
200. ANTTILA, Raimo: *Greek and Indo-European Etymology in Action. Proto-Indo-European *ág-*.2000.
201. DRESSLER, Wolfgang U., Oskar E. PFEIFFER, Markus PÖCHTRAGER and John R. RENNISON (eds.): *Morphological Analysis in Comparison*. 2000.
202. LECARME, Jacqueline, Jean LOWENSTAMM and Ur SHLONSKY (eds.): *Research in Afroasiatic Grammar. Papers from the Third conference on Afroasiatic Languages, Sophia Antipolis, 1996*. 2000.
203. NORRICK, Neal R.: *Conversational Narrative. Storytelling in everyday talk*. 2000.
204. DIRVEN, René, Bruce HAWKINS and Esra SANDIKCIOGLU (eds.): *Language and Ideology. Volume 1: cognitive theoretical approaches*. 2001.
205. DIRVEN, René, Roslyn FRANK and Cornelia ILIE (eds.): *Language and Ideology. Volume 2: cognitive descriptive approaches*. 2001.
206. FAWCETT, Robin: *A Theory of Syntax for Systemic-Functional Linguistics*. 2000.
207. SANZ, Montserrat: *Events and Predication. A new approach to syntactic processing in English and Spanish*. 2000.
208. ROBINSON, Orrin W.: *Whose German? The ach/ich alternation and related phenomena in 'standard' and 'colloquial'*. 2001.
209. KING, Ruth: *The Lexical Basis of Grammatical Borrowing. A Prince Edward Island French case study*. 2000.
210. DWORKIN, Steven N. and Dieter WANNER (eds.): *New Approaches to Old Problems. Issues in Romance historical linguistics*. 2000.
211. ELŠÍK, Viktor and Yaron MATRAS (eds.): *Grammatical Relations in Romani. The Noun Phrase*. 2000.
212. REPETTI, Lori (ed.): *Phonological Theory and the Dialects of Italy*. 2000.
213. SORNICOLA, Rosanna, Erich POPPE and Ariel SHISHA-HALEVY (eds.): *Stability, Variation and Change of Word-Order Patterns over Time*. 2000.

214. WEIGAND, Edda and Marcelo DASCAL (eds.): *Negotiation and Power in Dialogic Interaction.* 2001.
215. BRINTON, Laurel J.: *Historical Linguistics 1999. Selected papers from the 14th International Conference on Historical Linguistics, Vancouver, 9-13 August 1999.* 2001.
216. CAMPS, Joaquim and Caroline R. WILTSHIRE (eds.): *Romance Syntax, Semantics and L2 Acquisition. Selected papers from the 30th Linguistic Symposium on Romance Languages, Gainesville, Florida, February 2000.* 2001.
217. WILTSHIRE, Caroline R. and Joaquim CAMPS (eds.): *Romance Phonology and Variation. Selected papers from the 30th Linguistic Symposium on Romance Languages, Gainesville, Florida, February 2000.* 2002.
218. BENDJABALLAH, S., W.U. DRESSLER, O. PFEIFFER and M. VOEIKOVA (eds.): *Morphology 2000. Selected papers from the 9th Morphology Meeting, Vienna, 24–28 February 2000.* 2002.
219. ANDERSEN, Henning (ed.): *Actualization. Linguistic Change in Progress.* 2001.
220. SATTERFIELD, Teresa, Christina TORTORA and Diana CRESTI (eds.): *Current Issues in Romance Languages. Selected papers from the 29th Linguistic Symposium on Romance Languages (LSRL), Ann Arbor, 8-11 April 1999.* 2002.
221. D'HULST, Yves, Johan ROORYCK and Jan SCHROTEN (eds.): *Romance Languages and Linguistic Theory 1999. Selected papers from 'Going Romance' 1999, Leiden, 9-11 December.* 2001.
222. HERSCHENSOHN, Julia, Enrique MALLÉN and Karen ZAGONA (eds.): *Features and Interfaces in Romance. Essays in honor of Heles Contreras.* 2001.
223. FANEGO, Teresa, María José LÓPEZ-COUSO and Javier PÉREZ-GUERRA (eds.): *English Historical Syntax and Morphology. Selected papers from 11 ICEHL, Santiago de Compostela, 7-11 September 2000.* 2002.
224. FANEGO, Teresa, Belén MÉNDEZ-NAYA and Elena SEOANE (eds.): *Sounds, Words, Texts and Change. Selected papers from 11 ICEHL, Santiago de Compostela, 7-11 September 2000.* 2002
225. SHAHIN, Kimary N.: *Postvelar Harmony.* n.y.p.
226. LEVIN, Saul: *Semitic and Indo-European. Volume II: comparative morphology, syntax and phonetics; with observations on Afro-Asiatic.* 2002.
227. FAVA, Elisabetta (ed.): *Clinical Linguistics. Theory and applications in speech pathology and therapy.* 2002.
228. NEVIN, Bruce E. (ed.): *The Legacy of Zellig Harris. Language and information into the 21st century. Volume 1: philosophy of science, syntax and semantics.* 2002.
229. NEVIN, Bruce E. and Stephen JOHNSON (eds.): *The Legacy of Zellig Harris. Language and information into the 21st century. Volume 2: computability of language and computer applications.* 2002.
230. PARKINSON, Dilworth B. and Elabbas BENMAMOUN (eds.): *Perspectives on Arabic Linguistics XIII-XIV. Papers from the Thirteenth and Fourteenth Annual Symposia on Arabice Linguistics.* 2002.
231. CRAVENS, Thomas D.: *Comparative Historical Dialectology. Italo-Romance clues to Ibero-Romance sound change.* 2002.
232. BEYSSADE, Claire, Reineke BOK-BENNEMA, Frank DRIJKONINGEN and Paola MONACHESI (eds.): *Romance Languages and Linguistic Theory 2000. Selected papers from 'Going Romance' 2000, Utrecht, 30 November - 2 December.* 2002.
233. WEIJER, Jeroen van de, Vincent J. van HEUVEN and Harry van der HULST (eds.): *The Phonological Spectrum. Part I: Segmental structure.* n.y.p.
234. WEIJER, Jeroen van de, Vincent J. van HEUVEN and Harry van der HULST (eds.): *The Phonological Spectrum. Part II: Suprasegmental structure.* n.y.p.