Linear Unit Grammar

Studies in Corpus Linguistics

SCL focuses on the use of corpora throughout language study, the development of a quantitative approach to linguistics, the design and use of new tools for processing language texts, and the theoretical implications of a data-rich discipline.

Volume 25

Linear Unit Grammar: Integrating speech and writing
by John McH. Sinclair and Anna Mauranen

Linear Unit Grammar

Integrating speech and writing

John McH. Sinclair
Tuscan Word Centre

Anna Mauranen
University of Helsinki

John Benjamins Publishing Company
Amsterdam / Philadelphia

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

Cover design: Françoise Berserik
Cover illustration from original painting *Random Order*
by Lorenzo Pezzatini, Florence, 1996.

Library of Congress Cataloging-in-Publication Data

John McH. Sinclair
Linear Unit Grammar : Integrating speech and writing / John McH. Sinclair and Anna Mauranen.
p. cm. (Studies in Corpus Linguistics, ISSN 1388-0373 ; v. 25)
Includes bibliographical references and index.
1. Grammar, Comparative and general.
P151.S667 2006
415--dc22 2006051041
ISBN 978 90 272 2298 5 (Hb ; alk. paper)
ISBN 978 90 272 2299 2 (Pb ; alk. paper)
ISBN 978 90 272 9306 0 (Eb)

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

Table of contents

In Memoriam David Brazil 1925–1995

The influence of David Brazil's subtle and innovative thinking will be found in many places in this book, not only where we specifically cite him. For twenty years he led generations of students at The University of Birmingham to find their own feet in his chosen domains of language teaching, discourse intonation and, latterly, syntagmatic grammar. He died just after his last book, *A Grammar of Speech* was published, and before he could actively promote the refreshing novelty of his view of text as a series of increments rather than sentences. As we developed Linear Unit Grammar we felt ourselves moving ever closer to David's position, and we hope that this book may serve to rekindle interest in David's work.

For more about the life and work of David Brazil see:
http://www.speechinaction.net/SPARC_Brazil.htm

Acknowledgements

Our thanks are due to the many colleagues who helped this project by their participation in the two workshops that we conducted, at Jaio Tong University in Shanghai in October 2003 and in Ann Arbor, Michigan in May 2005. Special thanks go to Professors Yang Hui-Zhong and John Swales respectively for their congenial hosting of the events, and to ICAME/AAACL for including the second workshop in the programme of their annual conference.

We owe a debt of gratitude to all those involved in making the data available to us. Most of those whose conversations are transcribed remain by convention anonymous, but public-spirited none the less. The offer of a sample from the Hong Kong Corpus of Spoken English came at a crucial point in our work, and we were encouraged by the interest shown by Professors Martin Warren and Winnie Cheng throughout the last year of the project.

Preamble

You are encouraged to work through the following introductory activity as a preliminary to the presentation of Linear Unit Grammar. Please look at Figure 1.

theheadmasterofharrowtellsannmcferranwhyhehasl
etthetvcamerasintoaschoolfullofodditiesbarnabylen
on30thheadmasterofharrowschoolleansoverhisdesk
therearemoreimportantthingsinlifethanstrawboaters

Figure 1.

We expect that before you are consciously aware of having thought about what Figure 1 is, you will have formulated a number of hypotheses about it and will have decided on a number of tactics for dealing with it, and will be first of all aware of the preliminary results of all this involuntary activity. It is a written verbal communication. There are traces of English words in it. It is a string of characters without spaces or punctuation or even capital letters.

Perhaps almost immediately, given your linguistic training, competence in English and expectations of this book, you will guess that it is a piece of written English rendered solely as a sequence of characters. You are then able to formulate a stratagem for dealing with it, and you settle on the "traces of English words" that was one of your immediate reactions. You decide that you will attempt to turn it into a string of English words, and then see if it makes sense. Perhaps some investigation that you are conducting subliminally reports at this point that there are signs of coherent phrases here and there, strengthening your hypothesis.

Not everyone will respond in the same way to Figure 1, and we would be interested to hear from anyone who has a strikingly different experience. The two points that we hope most readers will agree on are (a) that most of your reaction is involuntary, and the hypotheses and strategies are formulated with little or no conscious intervention; (b) you begin, mentally, to add word spaces, capital letters and punctuation in order to make sense of the passage. There is no need to copy it out and add all these features, because once you have the idea it will become almost immediately readable. Occasionally you may have to backtrack or read a string of characters several times before the word spaces become obvious, because there are some specific difficulties in this passage — one of the reasons why it was chosen. The proper names are slightly unusual, there are some words and phrases that may not

be familiar to many readers — like *straw boaters* at the end. You may not be aware of Harrow school, an old English school, where hats called straw boaters are, or were, worn in the summer. On the helpful side there are repetitions like *the headmaster of harrow* and a cliché, *there are more important things in life than*, which, once noticed, explains almost a fifth of the whole passage.

The passage begins easily, because the most frequent word in English, *the*, heads the text. *Headmaster* is recognisable; in English it can be written in three different ways, *head master, headmaster* and *head-master*, and with or without initial capitals. In the original of this passage, curiously, the two instances are spelt differently, as a single word and as two separate words. *Of* is the second commonest word in English, fairly easy to spot; *harrow* may not be readily recognised, but *tells* is clear enough, so *harrow* must be something you can be headmaster of. If you saw the word *arrow* there you would have to revise your guess or have an *h* left over. The same goes for the *s* following *tell.* Perhaps at first its role as the present singular inflection is not obvious, and *san* could be the start of a name; the run of consonants *nnmcf* is a little off-putting, and indeed the whole string until *why* is encountered may be a puzzle. The clue is that many Scottish family names begin with *mc* or *mac*, signifying "son of". *McFerran* or *Mcferran* are plausible names, though unfamiliar to the authors; the original text reads *mcferran* which is certainly short of at least an initial capital. The occasional inconsistencies in the original text show that users of a language encounter routinely the same kinds of problems that the reader of Figure 1 is faced with, though usually diluted with plenty of unproblematic text. Once Ann Mcferran has been identified, the rest falls into place, and the first line is all but deciphered, because *he* and *has* are easy to spot. The line break splits *let*, but with *the* following it can be reconstructed quickly. *TV* is almost universal, and supported by *cameras*. By now the decipherer can polish off *into a school full of* without much perplexity, but the last few characters in the line may need a moment; *oddities* is not a common word, and *Barnaby* is not a common name, nor is *Lenon* unless you award it a double *n* in the middle and think of the Beatles.

The beginning of line 3 is distracting unless Mr Lenon has been identified, but 30^{th} need not detain us, and the next phrase has occurred before, here with the helpful word *school* following for those not so familiar with UK institutions. By now *leans over his desk* should be easy because, unconsciously, we have become trained in this medium even over such a short passage, and headmasters have desks anyway. The last line is a quote from Mr Lenon, which we have already construed. (This extract is the beginning of a piece in The Times of 10th June 2001).

The task of separating words in a piece of ordinary printed matter is an unfamiliar one for most readers, but one that we adapt to readily, presumably working out ad hoc tactics as we go along. The keys to efficient performance include:

(a) the ability to apply a hierarchical model to the linear string — in this case to postulate, correctly, that the passage consists of a string of word tokens, and that a placement of word boundaries will make the passage instantly legible and understandable.
(b) the ability to *prospect*, to look ahead for features that will help the interpretation of a difficult passage or settle a question of alternatives
(c) the ability to hold provisional interpretations in mind and to abandon them if they are superseded by more plausible ones — otherwise to continue with them perhaps without resolution — like how exactly to spell *mcferran*.

In the rest of this book we will be applying essentially the same techniques to a later stage of the process of "making sense of" text. We will not artificially hamper ourselves as we have done in this illustrative exercise, but will use the same arguments, rely on similar perceptions, knowledge and abilities in the reader, and chart the progress of text from its fairly raw state in a range of situations to something that makes reasonably good sense.

Introduction

This book is the report of a study of language in use and how people manage it, handle it, cope with it and interpret it. The main focus is on informal spoken English, but the structural statements are intended to apply to all varieties of English, whether written or spoken, whether standard or non-standard, whether specialised or general.

This is an unusual and rather bold claim. Most descriptions concentrate on one language variety, whether they say so or not[1], and the descriptions often perform poorly with any variety other than the one chosen. The vast majority of grammars concentrate on the structure of carefully written English; discourse analysis concentrates on semi-formal spoken language, conversation analysis on informal and intimate spoken varieties. There are also studies of specialised varieties, but the texts studied fall within one or other of the major varieties just mentioned; the contrastive study of different varieties also keeps constant such parameters as spoken vs. written, formality and preparedness[2].

It is central to our position that all varieties of a language in use can be described using the same descriptive apparatus, in contrast to the present state of affairs, where grammars make little effort to be flexible. Linear Unit Grammar is a descriptive apparatus and method which aims at integrating all or most of the superficially different varieties of English; it does not attempt to replicate the kind of analysis that received grammars perform, but organises the text into tractable units for further analysis, whether conventional or any more innovative analysis. Its main function is to show step by step how a latent hierarchy can be discerned in the linear string of word forms.

The underlying hypothesis that guides this stage is that a person applies essentially the same creative/interpretive apparatus to any language text, rather than that we have to postulate the existence of more than one such apparatus.

The aim of reconciling different language varieties was not the original motivation of the study. Like any open-ended investigation, it had aims and objectives which were modified by interim results and by the experience of undertaking the research. To begin with, the project was little more than an informal probe search-

---

1. With some recent exceptions such as Biber (1988), Biber et al. (1999) and Carter and McCarthy (2006).

2. There are exceptions, of course, such as MICASE (http://www.hti.umich.edu/m/micase/), which has collected examples of academic spoken English from a wide range of encounters on the scale of formality.

ing for answers to three questions:

(a) Is it possible to divide running text into chunks by assuming, and calling on, speakers' perceptions about the divisibility of text?
(b) If so, do speakers divide up the same text in similar ways?
(c) If so, does the result of making the divisions provide a foundation for a meaningful analysis of text?

It became clear that the answers to all these questions tended towards the positive, and this feedback gave rise to further questions

(d) what role do chunks play in the analysis of text (i.e. the description of the way in which meaning can be systematically derived from text)?
(e) what range of text types can be naturally chunked?

The answer to (d) was along the lines of "central and pivotal in the early stages", and the answer to (e) was that apparently any text type could be chunked.

At this point a plan of campaign began to take shape. We had begun with a text which showed no boundaries except word spaces — the minimal transcript of an informal spoken conversation. Our approach to analysis through chunking allowed us to take the early steps without regard for the details of the internal structure of the chunks, so we could cope with texts that contained "irregularities" of many kinds without the system breaking down. In a conventional analysis it is often impossible to proceed at one level if the text manifests an unusual pattern at a more detailed level, because all statements in such grammars are mutually dependent. In contrast, our initial request to assign provisional unit boundaries did not require any explicit awareness of the internal structure of the chunks or their combinations — it was only the boundaries that were asked for.

We realised that this property of our approach made it possible to investigate, or at least to begin an investigation into, texts and text types which were not normally amenable to structural description. One of us (Mauranen) had begun a long-term study of English as a Lingua Franca, and we found that a modified version of our original analytical system could make a satisfactory preliminary analysis of a lengthy sample of this variety, and also brought out distinctions between this and our first text which were intuitively satisfying (see Chapter 10).

The word "preliminary" above is important, because our analysis is not intended to reveal all the structural intricacies of the texts, but rather to bring out their similarities and differences in a systematic fashion, so that the output of our analysis should be close to an acceptable input into any of the established descriptive grammars. There is still a problem in the expectancy of the established grammars in that they do not tolerate even small divergences from their requirement of well-formedness, so we developed an extension into structural description that overlaps with

normal categories but is slightly more generalised, and so allows relationships to be determined while the data representation is still heavily linear (see Chapter 12).

In this we returned to the same stratagem that we started with — to concentrate not on the item, the token, the word, but on the relationship, the spaces in between. The inherent power and simplicity of this stratagem was demonstrated by Coniam (1998) in his "boundary" program devised over ten years ago, but it has not been followed up in software development. There is an implied claim in this stage of our work that the job of analysis is made easier and more accurate by separating the identification of units from the classification of constituents. When a text is divided into coherent units, and within those units the basic structural relationships have been assigned, the job of relating the constituent items to classes should be much easier than starting from scratch. This is a similar strategy to a "bootstrap" operation in computing, where initially the system can only cope with superficial and unvarying entities, but passes through stages of ever-increasing sophistication. In our case, it is guided by coming ever closer to embodying the meaning.

We then decided on several further text exemplars. Because of the novelty of our analysis we had to choose brief extracts, but increased our range of variety. We asked a team of researchers in The Polytechnic University of Hong Kong who were building a corpus of local spoken English (HKCSE) to send us a transcript of their choice — but not to include the sound recording.

Around this time we began to consider how adaptable the emerging system of analysis might be with respect to written varieties of English, so we decided to select three short passages and apply the same procedure to them as we had used in the spoken transcripts. One started from an earlier, unsatisfactory attempt to describe "compressed" English (Sinclair 1988), the highly abbreviated language associated with some reference books, and in the present era, text messages; the second was a passage from a famous novel (Joyce 1922) where the author seemed to be writing a simulation of speech; finally we chose a piece of written English which was as representative of ordinary expository prose as we could find — an editorial from *The Independent* newspaper.

We were pleased to find that our system coped with these completely different texts with only minor adaptation, and that the comparisons that we made after analysis continued to be intuitively satisfying, so we decided that we should publish our results in order to get feedback from the academic community.

Existing grammars

From time to time we make reference to the prevailing state of grammatical knowledge, the aims of grammar construction and the styles of presentation of

grammatical information in a wide range of influential publications. Usually the context is the contrast between our approach and the main lines of what has gone before. To make the contrasts, we do not usually need to pick out one or more specific examples of received, conventional, etc. grammars, but we just lump them together. While this is rough justice, since each of them, and each category of them, has its own position in the panoply, we need to make it constantly clear that it is not our intention to mount specific criticisms of them; simply to make contrasts.

On the whole we expect readers to have sufficient familiarity with existing grammars to appreciate that our references are general rather than vague. But here is a brief checklist of the features that characterise Linear Unit Grammar, and which are not prominent in any other grammars:

(a) the maintenance of linearity in the description wherever possible
(b) the syntagmatic orientation of the description (in contrast with the paradigmatic orientation of most grammars)
(c) the "bottom-up" approach to description, though mediated heavily by intuition from the very first step
(d) the cyclical, "bootstrap" style of analysis as against the description of sentences in a single pass through the grammar
(e) the acceptance of any alphanumeric string that has good reason to be considered an instance of English text (in contrast with the basis of most grammars on the written form of the language).

This checklist anticipates much of the book that is to come, but may be useful to keep in mind.

Structure of the book

The reader's very first experience of this approach will be, we hope, the Preamble. There, in a very small exercise, a taster for the kind of data and argument is offered. After this Introduction we turn to the book proper, which is organised into three sections.

The first section is entitled *Preliminaries*. Chapter 1 raises the main conceptual points that the argument relies on, and illustrates by one brief and one extended example the general principles of analysis. Chapter 2 traces the origins of these conceptual orientations in previous literature, and attempts to acknowledge the important influences on our thinking. The early part deals with previous grammatical research, later on it tackles key works in psychology, education and applied linguistics. Chapter 3 is a short account of the textual data that we chose to describe.

The second section, *Analysis*, goes through the analytic method, which we call

Linear Unit Grammar (LUG) on a step-by-step basis which is much more than a procedure. After an overview of the descriptive method in Chapter 4, there are five chapters each of which describes a step in the analytic process. Chapter 10 rounds off this section with a short commentary on the six example passages in the light of the descriptions.

The third section raises, in Chapter 11, matters of theory on a broader basis than hitherto, using the experience of analysis as a guide. Chapter 12 rounds off the book by considering some developments and applications that interest us. First it considers the consequences of continuing the LUG kind of description into the area of parsing, where it begins to offer alternatives to conventional grammars; then it assesses the possibilities for automation, and finally the implications for Applied Linguistics.

The book finishes with an Appendix which reports on one of the Workshops that shaped the book, which was generously hosted by ICAME/AACL at Ann Arbor in May 2005.

This book has been written jointly, and each part of it has been discussed extensively. The core of the book, the analyses, were constructed in a lengthy process mainly using e-mail. This required each of the authors to make an individual analysis, step by step, without knowledge of the other's decisions, and for the discrepancies to be resolved by exchanging documents, often many times, finally reaching conclusions at one of several meetings we were able to have despite the physical distance between Helsinki and Florence.

The responsibility, therefore, is fully shared, but it is normal in such a publication to indicate which of us initiated the drafts of the chapters. AM drafted Chapters 3, 6 and 8, the commentaries on *ELFA*, *HKCSE* and *Independent* texts in Chapter 10, and the Appendix. Chapter 2, placing the ideas in their cultural roots, and Chapter 12, the applications, were written half-in-half, while Chapter 9, the final step in analysis, was a completely joint effort. JS drafted the rest, i.e. the Preamble, Preface, Chapters 1, 4, 5, 7, the *Lexis*, *Gazetteer* and *Joyce* texts in Chapter 10, and Chapter 11.

Terms and concepts

This book contains detailed analysis and discussion of the structural detail of texts, and we have tried to keep novel terminology to a minimum, though some is inescapable. The three terms that distinguish the study are:

Linear Unit Grammar (LUG). This is the name we have chosen for this descriptive model. It is a grammar expressed as far as possible in a linear succession of units.

Provisional Unit Boundary (PUB). The first step in analysis is the division of the text into chunks, which we separate with boundaries and call them PUBs.

Linear Unit of Meaning (LUM). After the chunks are classified they are recombined into units, now directly meaningful, called LUMs.

We do not define a chunk because we are using it as a pre-theoretical term. But on several occasions in the explanations we discuss the notion of a chunk in order to be as clear as possible about how we are using the word.

So our first step in analysing a text is to divide it into alternating chunks and boundaries (*PUB*s). Then we classify the chunks, and call them *elements*. Then we combine the elements into finalised meaningful segments which we then call *LUM*s.

Other words like "segment", "fragment" are used informally from time to time when we do not want to be more specific (but note that our term for the MF element is *message fragment*). Throughout, we refer to places in our transcripts as *lines* with a number; this is for reference purposes only, and the numbering in each example is particular to the example in its particular place in our argument.

Throughout, we use *text* or *texts* irrespective of whether we are referring to transcripts of a spoken encounter or written documents. We require a single term to talk about all our data, and indeed any sequence of alphanumeric characters, with or without punctuation. It is also convenient to use a term which is heavily associated with written language even though we are mainly talking about spoken varieties; our data consists entirely of material in written form, and we deliberately do not invoke aspects of spoken performance, even if we have access to recordings, so that the reader can follow our arguments and decisions directly with the data.

The term *message* is one of our main structural labels, and we have considered and reconsidered the term, because the concept it labels is open to various interpretations, even misunderstandings, and needs careful definition. We want a simple and transparent term, but in this area all the available words are open to misunderstandings; *topic, subject, subject matter, shared knowledge, shared experience* — none of these is "safe". We decided to pick our way in this minefield rather than devise special terms which would not be readily accessible to the reader. With proper safeguards, most of the terms listed above can be used.

In the case of *message* there are two possible inferences that we want to avoid:

(a) a message element is not some meaning which is coded into speech or writing and then decoded by the listener or reader; its meaning is integral to the way in which it is expressed

(b) message elements are not the only carriers of meaning; meanings which depend on the circumstances of real-time interaction are not expressed in textual message units, and so in LUG are paraphrased in parallel descriptive notes.

Message elements combine into message units. A message unit is a coherent stretch of text whose meaning is interpreted according to the structural conventions of the language. Its purpose is to update the virtual world of shared experience of the participants in the spoken or written interaction by means of topic incrementation. For the notion of *increment* we rely on Brazil's (1995) work, explained in Chapter 2.

The shorthand labels for the analytical categories are named below:

O	organisational element
OI	interactive organisational element
OT	text-oriented organisational element
M	message-oriented element
MF	message fragment
M–	incomplete message unit
+M	completion of message unit
+M–	partial completion of message unit
MS	supplement to message unit
MA	adjustment to message unit
MR	revision to message unit

Note that each LUM contains one, and only one, M.