

Edited by:
Peter Juel Henriksen

Linguistic Theory and **Raw Sound**

40

Copenhagen Studies in Language

Linguistic Theory and Raw Sound

Linguistic Theory and Raw Sound

Edited by:
Peter Juel Henriksen

Copenhagen Studies in Language 40

Samfundslitteratur

Linguistic Theory and Raw Sound

Copenhagen Studies in Language 40

© Samfundslitteratur and the authors

ISSN 0905-09857

ISBN 978-87-593-9753-4

1. e-edition 2011

Cover by SL Grafik

Layout by the editor

Copenhagen Business School

Published by:

Samfundslitteratur Press

Rosenørns Allé 9

DK-1970 Frederiksberg

Phone: + 45 38 15 38 80

Fax: + 45 35 35 78 22

sforlagene@samfundslitteratur.dk

www.samfundslitteratur.dk

Editor:

Peter Juel Henriksen

Contents

| | |
|---|-----|
| Editor's Preface | 7 |
| Words and alternative basic units for linguistic analysis | 9 |
| <i>Jens Allwood, A. P. Hendrikse and Elisabeth Ahlsén</i> | |
| Den Lilla Trekant - Learning Danish shape and color terms from scratch | 27 |
| <i>Peter Juel Henriksen</i> | |
| Stress, pauses, pronominal types and pronominal functions in Danish spoken data | 45 |
| <i>Costanza Navarretta</i> | |
| Looking for lost syllables in Danish spontaneous speech..... | 61 |
| <i>Ruben Schachtenhaufen</i> | |
| New directions for the Uniqueness Point..... | 87 |
| <i>Laura Winther Balling</i> | |
| Exploring classification of Swedish dialects by speech melody | 101 |
| <i>Marcus Uneson</i> | |
| Frequency Selective Filtering of the Modulation Spectrum and its Impact on Consonant Ident..... | 119 |
| <i>Thomas U. Christiansen and Steven Greenberg</i> | |
| What is identified and what is lost in aphasia? | 135 |
| <i>Elisabeth Ahlsén</i> | |
| Information structure and communicative functions in spoken and multimodal data | 149 |
| <i>Patrizia Paggio and Philip Diderichsen</i> | |
| Ask the experts - Part I: Elicitation..... | 169 |
| <i>Jens Edlund and Joakim Gustafson</i> | |
| Ask the experts - Part II: Analysis | 183 |
| <i>Jens Edlund and Joakim Gustafson</i> | |
| An Audio-Visual Approach to Measuring Discourse Synchrony in Multimodal Conversation Data | 199 |
| <i>Nick Campbell</i> | |
| On the many benefits of a multidimensional approach to the analysis of spoken dialogue | 209 |
| <i>Harry Bunt and Volha Petukhova</i> | |

The essence of language is human activity – activity on the part of one individual to make himself understood by another, and activity on the part of that other to understand what was in the mind of the first. These two individuals, the producer and the recipient of language, or as we may more conveniently call them, the speaker and the hearer, and their relations to one another, should never be lost sight of if we want to understand the nature of language and of that part of language which is dealt with in grammar. But in former times this was often overlooked, and words and forms were often treated as if they were things or natural objects with an existence of their own – a conception which may have been to a great extent fostered through a too exclusive preoccupation with written or printed words, but which is fundamentally false, as will easily be seen with a little reflection.

Otto Jespersen

The Philosophy of Grammar

Editor's Preface

"Linguistic Theory and Raw Sound 2009" – this was the title of the *con amore* symposium held at Marston Hill, Mullsjö, Sweden, in the late summer of 2009. The call for participation was a single motivating question.

What kind of linguistic information
can be distilled from recordings of spontaneous speech
without presupposing any, or much, theoretical knowledge?

Linguistic theory is committed to describing and explaining *how* and *why* human language users communicate. Theory must therefore take its point of departure in data collected in the living world, from the midst of the communication situations. Raw data. Sometimes we have to gather data from arranged situations, having informants solve particular language tasks in a lab, or we must even resort to fossilized language samples, such as written text, in order to simplify our theoretical groundwork. We should however never forget Jespersen's warnings to the aspiring grammarian (even if he perhaps did not always remember himself). Words and sounds are born as carriers of meaning and tools for understanding. This is why we met in Mullsjö, empiricists and formalists side by side. To let raw data and sophisticated theory cross-fertilise in creative debate and daring new ideas.

This book brings forward some of the papers presented at the symposium. The papers are organized in three main sections, the first section covering the **meaningful units of language** and how to identify them. The second section is on **production and perception** of speech, while the final section deals with aspects of **dialogue and multi-modal communication**.

CMOL

CMOL (Center for Computational Modelling of Language) is supported by the Danish Council for Independent Research (Det Frie Forskningsråd, FKK, <http://www.fi.dk>). Research at CMOL is focused on creating formal models of human linguistic performance. *Spoken language studies* is a corner stone of the research programme and include:

- models of speech perception
- models of first language acquisition
- comparative studies of spoken languages (Nordic and World)
- tools for computer assisted language learning (CALL)

Editor's note

The style sheet and other requirements of Copenhagen Studies in Language forced us to revise the graphical layout of the original papers. In spite of all good intentions, this may have introduced some errors in the process. Any such should be blamed on the editor, not on the authors.

I wish to thank cand.mag. Jeppe Beck for his extensive and valuable work with the textual layout of this publication.

Extra copies of this publication can be obtained from CMOL. Please contact the editor.

Peter Juel Henriksen
pjh.isv@cbs.dk

CMOL (ISV)
Copenhagen Business School
Dalgas Have 15
DK-2000 Copenhagen

Words and alternative basic units for linguistic analysis

Jens Allwood

SCCIIIL Interdisciplinary Center, University of Gothenburg

jens@ling.gu.se

A. P. Hendrikse,

Department of Linguistics, University of South Africa, Pretoria

hendrap@unisa.ac.za

Elisabeth Ahlsén

SCCIIIL Interdisciplinary Center, University of Gothenburg

elisa@ling.gu.se

Abstract

The paper deals with words and possible alternative to words as basic units in linguistic theory, especially in interlinguistic comparison and corpus linguistics. A number of ways of defining the word are discussed and related to the analysis of linguistic corpora and to interlinguistic comparisons between corpora of spoken interaction. Problems associated with words as the basic units and alternatives to the traditional notion of word as a basis for corpus analysis and linguistic comparisons are presented and discussed.

1. What is a word?

To some extent, there is an unclear view of what counts as a linguistic word, generally, and in different language types. This paper is an attempt to examine various construals of the concept “word”, in order to see how “words” might best be made use of as units of linguistic comparison. Using intuition, we might say that a word is a basic linguistic unit that is constituted by a combination of content (meaning) and expression, where the expression can be phonetic, orthographic or gestural (deaf sign language). On closer examination, however, it turns out that the notion “word” can be analyzed and specified in several different ways. Below we will consider the following three main ways of trying to analyze and define what a word is:

- (i) Analysis and definitions building on observation and supposed easy discovery
- (ii) Analysis and definitions building on manipulability
- (iii) Analysis and definitions building on abstraction

2. Analysis and definitions building on observation and supposed easy discovery

We will start by considering analyses and definitions intended to build on observation of linguistic communication. Here the idea is that words are the basic building blocks of linguistic communication, providing combinable units of meaning and external expression that should as such be fairly directly observable and discoverable when inspecting linguistic communication, whether in written, spoken or gestural form. This can especially be seen in the definition of orthographic words given below.

(i) Orthographic words

According to Trask (2004) “[a]n orthographic word is a written sequence which has a white space at each end but no white space in the middle”.

This definition of “orthographic word” is both too wide and too narrow, in relation to other notions of word that intuitively have precedence. For example, the expression “rail road” has two orthographic words but is intuitively one word. This means that the notion of orthographic word as defined captures too much, i.e. it is too wide => too many words.

But making “rail road” into two orthographic words is also too narrow => not capturing the word (semantic unit and phonological stress unit, lexeme) that is actually there.

(ii) Phonological words

Following Trask again, Trask (2004) defines a phonological word as “a piece of speech which behaves as a unit of pronunciation according to criteria which vary from language to language”

Unfortunately, there are units other than words that perhaps meet such phonological requirements, for example phonemes, syllables or breath groups. The definition does not tell us how to differentiate these units from each other. The mention of language-specific features does not help, since, these might be different for different languages, for the different units. In addition, when transcribing words, i.e. making them into orthographic words, typical phonetic information that may be used in the identification of phonological words such as stress, tone patterns, pauses (length) are either typically not represented in transcriptions. One reason for this is that such information is not traditionally part of written language, another is that it may reflect that this information is not so easily consciously recognized/observed by phonetically untrained transcribers.

If we consider the relation between orthographic words and phonological words, we first may note that given these two definitions of a word, a consequence is that there is no 1- 1 correspondence between orthographic words, phonological and semantic words. Consider the following examples: *rail road* (2 orthographic words - 1 phonological word) or *I'm*, *you're*, *won't* and *ain't* (1 orthographic word - 1 phonological word but two semantically motivated words). *New York* (2 orthographic words) vs. *Newfoundland* (1 orthographic word). *New York and Newfoundland*, thus, fairly arbitrarily, have different orthographic status while both probably are single phonological words etc.

(iii) **Gestural words**

Using Trask's definition of phonological words as a model, we can now define gestural words analogously as as "a piece of gestural communication which behaves as a unit of gesturing according to criteria which vary from language to language"

The relation between orthographic, phonological and semantically motivated and gestural words is more complex, so that 1 – 1 correspondences between the three word forms are not always possible to establish here either. Concerning gestural languages (sign languages), one reason for this is that while written and spoken words can be seen as

variants of the same unit in two different expressive modes, gestural words in sign language are units in a new language and not gestural variants of the same word, in the sense that the written and spoken variants of a word are variants.

We can also note that only the definition of orthographic word is operational, i.e. lives up to the desiderata of being both directly observable and discoverable and thus, directly usable as an element in automated information retrieval. Since, as we have seen above, the criteria given for what is a unit of pronunciation or gesturing are not sufficient, these concepts thus remain in need of further specification and clarification.

3. Analysis and definitions building on manipulability

Many linguists have thought that word criteria, based on inherent word features that are supposed to be directly observable are unreliable and need to be supplemented by other criteria. Some widely used such criteria are criteria that in a syntactic mode focus on the unit status of words. Two criteria are often suggested:

- (i) Moveability
- (ii) Resistance to intrusion and interruption

Both of these criteria have often been used to define the notion “word”. We will now consider them one by one.

3.1 Moveability

According to this criterion, a word is the smallest element of a sentence that can be moved around without destroying the grammaticality of the sentence.

Thus, the fact that the word *often* in the expression *often he went to the house* can be moved from first to last position as in *he went to the house often*, shows that *often* is a word. A problem with this criterion is that