The London-Lund Corpus of Spoken English: Description and research Edited by Jan Svartvik (review)

W. Nelson Francis

Language, Volume 68, Number 1, March 1992, pp. 196-199 (Review)

Published by Linguistic Society of America
DOI: https://doi.org/10.1353/lan.1992.0072

——. 1975. Austro-Thai language and culture. New Haven: Human Relations Area Files Press.

——. 1985. Toppakō: Tōnan Ajia no gengo kara Nihongo e. Hi no kami no tami no kigen [A breakthrough: From the languages of East Asia to Japanese. The origin of the people of the sun god]. Translated by Nishi Yoshio. Tokyo: Project on Lexicographical Analysis, National Inter-University Research Institute of Asian & African Languages & Cultures.

Edmondson, Jerold A., and David B. Solnit (eds.) 1988. Comparative Kadai: Linguistic studies beyond Tai. Dallas: Summer Institute of Linguistics & University of Texas at Arlington.

——, ——. 1988. Introduction. In Edmondson & Solnit, 1–17.

Edmondson, Jerold A., and Yang Quan. 1988. Word-initial preconsonants and the history of Kam-Sui resonant initials and tones. In Edmondson & Solnit, 143–66.

Haudricourt, André-Georges. 1956. De la restitution des initiales dans les langages monosyllabiques: Le problème du thai commun. Bulletin de la Société de Linguistique de Paris 52.307–22.

——. 1967. La langue lakkia. Bulletin de la Société de Linguistique de Paris 62.165–82.

——. 1975. À propos du puzzle de W. J. Gedney. Studies in Tai linguistics in honor of William J. Gedney, ed. by Jimmy G. Harris and James R. Chamberlain, 252–8. Bangkok: Central Institute of English Language, Office of State Universities.

Maddieson, Ian. 1984. Patterns of sounds. Cambridge: Cambridge University Press.

Miller, R. A. 1987. Review of Benedict 1985. Lg. 63.643–8.

Shibatani, Masayoshi. 1990. The languages of Japan. Cambridge: Cambridge University Press.

Starostin, S. A. 1986. Problema genetičeskoi obščnosti altaiskikh jazykov. Istoriko-kul'turnye kontakty narodov altaiskoi jazykovoi obščnosti. Tezisy dokladow XXIX sessii Postoiannoi Meždunarodnoi Altaističeskoi konferencii [PIAC], v. 2, 94–112. Moscow.

Thomason, Sarah Grey, and Terrence Kaufman. 1988. Language contact, creolization, and genetic linguistics. Berkeley & Los Angeles: University of California Press.

Thompson, Laurence C. 1976. Proto-Viet-Muong phonology. Austroasiatic Studies, vol. 2, ed. by Philip N. Jenner, Laurence C. Thompson, and Stanley Starosta, 1113–204. Honolulu: University of Hawaii Press.

Department of Asian Languages
and Cultures
University of Michigan
Ann Arbor, MI 48109-1285

[Received 24 May 1991.]

The London-Lund Corpus of Spoken English: Description and research. Edited by Jan Svartvik. (Lund studies in English 82.) Lund: Lund University Press, 1990. Pp. 350.

Reviewed by W. Nelson Francis, *Brown University*

The thirteen papers which comprise this volume are all concerned in one way or another with the London-Lund Corpus of Spoken English (LLC). The dominant theme is the development of a realistic text-to-speech computer program—one in which written or printed English is converted into a naturalistic counterpart of spoken English. Such a program must produce not only the correct sequence of phonetic realizations of phonemes—which by itself produces the monotonous type of talk attributed to robots in popular films—but also a natural intonation and segmentation into tone units, often separated by various lengths of pauses with or without the fillers that carry on vocalism across a gap in the running context of meaningful speech. The relationship between syntactic-semantic structure, including punctuation, and phonic structure is by no means a simple one; many aspects of this relationship are addressed in the book under review.

The opening chapter (11–59), by Svartvik, the head of the LLC projects, and Sidney Greenbaum, the present director of the Survey of English Usage (SEU) on which these projects depend, gives a short description and history of the LLC and an inventory of its contents. This corpus comprises the spoken half of SEU, the million-word corpus of educated English collected at University College London under the direction of Sir Randolph Quirk, the former Quain Professor there. The entire corpus contains 200 texts of 5000 words each, half of which were recorded from various types of spoken English, some of it public (e.g. radio programs) and some of it private. The latter was in part recorded surreptitiously, a practice disallowed in the United States in spite of its value for capturing speech at its most natural and least self-conscious. The corpus was transcribed in London in standard orthography but using a rather elaborate system for recording intonation (see Crystal 1969). The LLC transfers this transcription to a database, with some simplification of the intonation marking.

Svartvik & Greenbaum list the 100 samples constituting the LLC with their participants, identified by sex, age, and occupation or status, but not by name. Since the collecting was done primarily at University College, it is understandable that a large proportion of the participants are classed as academics. Given the aim of SEU—to present a sample of 'the whole range of educated English usage, from learned and technical writing to the most spontaneous colloquial English' (Quirk 1968:80)—this does not result in noticeable stylistic skewing, though it undoubtedly affects the lexicon. This chapter concludes with a 12-page bibliography of publications using Survey material.

The larger part of the book brings together twelve papers under the heading 'Research'. These all relate to the project Text Segmentation for Speech (TESS), which has been the active center for computational work at Lund over the last several years. The papers are by four authors: Mats Eeg-Olofsson, Bengt Altenberg, Anna-Brita Stenström, and Svartvik himself.

Eeg-Olofsson is the computer specialist of the group and has done the programming for LLC and for TESS. In 'An automatic word class tagger and phrase parser' (107–36) he first describes a new tagging program, based on those developed for tagging the Lancaster-Oslo/Bergen (LOB) corpus, but more detailed and specific. The new tag set contains over 200 items, compared with the 83 used for Brown (Francis & Kučera 1982) and 132 for LOB (Johansson et al. 1986). Eeg-Olofsson lists the new set and describes the problems encountered in using it on a selection of three samples from the Brown corpus. He claims a success rate varying from 94 to 97 percent and feels that this cannot be much improved by using a more refined program. The greater part of this paper is devoted to description and exemplification of the automatic phrase parser developed in TESS.

This leads to the topic of segmentation of spoken language, which is the concern of most of the remaining papers. Three papers by Stenström take up in detail the 'discourse elements' in speech which have various functions in discourse—specifically, special lexical items, pauses, and 'adverbial commas'. Stenström's 'Lexical items peculiar to spoken discourse' (137–75) includes a set of single words and short phrases used in conversation to express agreement, encourage continuation, prevent interruption, and perform other functions aimed at the orderly progress of a conversation. Her discussion is clear and interesting and deals in considerable detail with a linguistic apparatus that has not been much discussed. She lists 67 examples, from *all right* to *I beg your pardon*, and classifies them into 16 groups on the basis of their function in interactive discourse. In her second paper, 'Pauses in monologue and dialogue' (211–53), Stenström divides pauses into three types: silent pauses (SP); filled pauses (FP) such as [ə:] and [ə:m]; and verbal fillers (VF), including phrases like *I mean* and *you know*, which create a pause in the syntax while continuing the phonic stream. She describes the positions in discourse where these appear, their relation to syntactic-semantic subdivisions, and their indication by punctuation in written texts.

'Adverbial commas' (254–66) is Stenström's term for sentential adverbials such as *frankly, obviously, indeed*, and even the much-maligned *hopefully*. She discusses the positions which these may fill and their relation to tone units on the one hand and syntactic structure on the other. Unlike pause fillers, which are unique to spoken discourse, these appear in written discourse as well, where they are often marked by comma punctuation. Stenström presents a list of 52 of these (256–7) with their tags and frequencies in the LOB and LLC corpora. She also gives a table of prosodic realizations, based on study of readings aloud from printed texts. Her conclusion is that there is not always a strict relationship between punctuation and intonation, so her rules must be taken as 'a compromise, influenced not only by previous research but also by linguistic intuition and common sense' (264).

Altenberg contributes an interesting paper on 'Spoken English and the dictionary' (177–92) in which he deals with the shortcomings of even the best modern dictionaries in supplying information about spoken usage. He illustrates his point by a discussion of the *Longman dictionary of contemporary English* (*LDOCE*) and the COBUILD dictionary of John Sinclair and his Birmingham colleagues. He concludes (189–90):

> 'Speech differs from writing in many fundamental ways. I have here touched on two speech-specific phenomena, the use of intonation to differentiate adverbial functions, and the use of lexical items with pragmatic functions that are difficult to describe in traditional grammatical terms. If we wish dictionaries to reflect the spoken language (which they surely should do), they must also recognize these phenomena and find methods of representing them in an adequate way.'

In two other papers, 'Predicting text segmentation into tone units' (275–86) and 'Automatic text segmentation into tone units' (287–323), Altenberg discusses a difficult problem. He develops eleven rules indicating where tone-unit boundaries are to be inserted into written discourse that is being converted to artificial speech. Some of these are quite complex, with strings of tags indicating tone-unit boundaries, often qualified by special conditions of the type and length of phrasal units. One of the simpler rules is paraphrased as follows: 'Rule 9 separates a preverbal subject from the rest of the clause, provided the subject consists of at least three words and there is a certain distance to the preceding higher-level boundary. The rule is blocked for passive clauses, which tend to be divided after the verb rather than the subject' (302).

Altenberg does not claim that his rules are part of the competence of native speakers, though that is certainly implied. A rather generous allowance claims 85% accuracy; he discusses reasons for the 15% of failures. Since his evidence is limited to ten of the 100 samples in LLC, there is a possibility that idiosyncratic variation among speakers may have influenced his rules. Furthermore, his analysis of 'correct' intonation patterns is based on the transcription of the texts included in the SEU. Perhaps because of differences between British and American intonation patterns, this reviewer had difficulty in accepting many of the 'correct' transcriptions and hence the terminology of some of the rules. But it must be granted that Altenberg's rules go a long way toward setting a realistic intonation pattern for artificial speech.

There is too little space here to comment on the remaining three papers—one by Altenberg on functions of the 'booster' (British for sudden pitch rise; 193–209), one by Eeg-Olofsson detailing the computer program used in text segmentation (325–36), and a brief description by Svartvik of a program in graphic English prosody developed with the aid of students in computer science (267–74). All three contribute in one way or another to the general theme of the book.

In sum, this book presents further evidence of the vigor and originality of computer corpus linguistics in Sweden. If ever we succeed in getting a program that will make computers talk like real people, this work will have made a major contribution to that goal.

REFERENCES

Crystal, David. 1969. Prosodic systems and intonation in English. Cambridge: Cambridge University Press.

Francis, W. Nelson, and Henry Kučera. 1982. Frequency analysis of English usage: Lexicon and grammar. Boston: Houghton Mifflin.

Johansson, Stig; Eric Atwell; Roger Garside; and Geoffrey Leech. 1986. The tagged LOB corpus: Users' manual. Bergen: Norwegian Computer Centre for the Humanities.

Quirk, Randolph. 1968. The Survey of English Usage. Essays on the English language, medieval and modern, 70–87. Bloomington: Indiana University Press.

Department of Cognitive and
Linguistic Sciences
Brown University
Providence, RI 02912

[Received 15 July 1991.]