LINGUISTICA

SERIES C

RELATIONES, 4.

ROBERT ILSON

ASSEMBLING, ANALYSING AND USING A CORPUS OF AUTHENTIC LANGUAGE

(A lecture given on the Survey of English Usage
at the Linguistics Institute of the
Hungarian Academy of Sciences
on 2 September, 1988)

A MAGYAR TUDOMÁNYOS AKADÉMIA NYELVTUDOMÁNYI INTÉZETE
INSTITUTUM LINGUISTICUM ACADEMIAE SCIENTIARUM HUNGARICAE

1991

LINGUISTICA
SERIES C
RELATIONES, 4.

ASSEMBLING, ANALYSING AND USING A CORPUS OF AUTHENTIC LANGUAGE

LINGUISTICA

SERIES C

RELATIONES, 4.

ROBERT ILSON

ASSEMBLING, ANALYSING AND USING A CORPUS OF AUTHENTIC LANGUAGE

(A lecture given on the Survey of English Usage
at the Linguistics Institute of the
Hungarian Academy of Sciences
on 2 September, 1988)

Transcribed and edited by

ANDREA ÁGNES REMÉNYI

A MAGYAR TUDOMÁNYOS AKADÉMIA NYELVTUDOMÁNYI INTÉZETE
INSTITUTUM LINGUISTICUM ACADEMIAE SCIENTIARUM HUNGARICAE

1991

Publication of this book has been made possible in part by
Országos Tudományos Kutatási Alap,
Élőnyelvi vizsgálatok, No. 3220.

ISBN 963 8461 58 6
ISSN 0866-4196

Prefatory Note

In reproducing Dr Robert Ilson's lecture given to an audience in Budapest about the Survey of English Usage, two guiding principles were used. First, an attempt was made to transcribe the tape-recorded lecture in a near-verbatim fashion but with an eye to readability. Second, the text which follows was not meant to be the demonstration of a method of transcription. It was meant to be a slightly edited version of the spoken lecture retaining, for instance, a good deal of false starts and hesitations.

The following conventions are used:

(1) No subtitles are given. Paragraphs and punctuation marks were established by the transcriber. An attempt has been made to follow English orthography. Intonation is not transcribed.

(2) False starts (interrupted words and structures) and slips of the tongue are preserved. Interrupted words end in a hyphen and interrupted structures are followed by a comma.

(3) Numbers are written as figures. When spelt out, abbreviations are written in uppercase letters, e.g. *K-W-I-C concordance*, *S text*.

(4) Contractions are indicated in the transcription.

(5) Pauses, filled pauses and lengthening as hesitation are not transcribed.

(6) Following Hartvig Dahl's *Word Frequencies of Spoken American English* (Essex, Connecticut: Verbatim, 1979), if a word was unclear and could not be discriminated by repeated listening to the tape, it was represented as Z. The number of Z's used is an estimate of the number of such words.

(7) The lecturer's and the audience's extralingual acts are enclosed in square brackets. So are the explanations of uncorrected slips of the tongue.

(8) Parts of the text in which the transcriber was uncertain are enclosed in double parentheses, e.g. ((a)).

The transcription of the recorded lecture has been checked by Dr Ilson.

Andrea Ágnes Reményi

This [points to the mike on his lapel] is a device that we have never used at the Survey of English Usage, probably because when the Survey of English Usage was founded in about 1959, I'm not sure that these things exist[ed], the, which I believe are called lavaliere microphones. And it certainly strikes me that there is a whole range of texts that we could have, and that indeed you could have, at the Survey of English Usage and similar projects, which would be made possible by this simple advance in the means of production of texts. So [laughter because the mike falls off his lapel]. That illustrates some of the problems of collecting spoken language.

Well, the official title of my talk is Assembling, Analysing and Using an Authentic Corpus, a Corpus of Authentic Language. And I believe that there is now in progress at this very institution a project of a similar nature, not by any means identical, but of a similar nature to the project for English that has been being carried out for more than 25 years as the Survey of English Usage at University College London. And it's about our experiences at the Survey of English Usage that I'd like to talk to you in the hope that some of what we have experienced will be of use to you in your projects.

So what is the Survey of English Usage?

The Survey of English Usage at University College is a corpus, and it serves as a data base to a concordance of various features of contemporary standard British English. And a lot of those notions: corpus, data base, concordance, various features of contemporary standard British English, a lot of those simple-sounding words need a bit of explanation.

In assembling this corpus there were a number of choices that we had to make at the very beginning of the enterprise. First of all, we had to determine the size, the overall size of the corpus. Then, we had to decide what types of text we were going to include

in the corpus. Then, we had to decide how long each text would be. We also had to decide what sources we would use f-, as legitimate for our texts and what sources of texts, what speakers, what writers and so on we would exclude, and what we would include. And having answered all those questions we then had to decide what categories of analysis we would use on the texts we had so laboriously assembled.

But one choice, one decision was already decided for us at the outset, and that was this: because the Survey was intended to lead to a concordance, as our modern computer-friends call it, a concordance of various features of the English language, we, it was already decided that we would have to take every token of every type that we decided to investigate. In other words, once we decided that we would take a certain category f-, o-, of information, for example, sequences of adjectives, or noun phrases, or verb phrases, or whatever, once we decided to investigate a particular category we were committed to taking every example of that category that turned up in our texts. That, in fact, is what I mean by saying that the Survey is a concordance, ((a)) complete collection of every example of certain categories which we believe to be relevant to the study of English grammar.

Now there's one, before going into more details about what this all means, there is one important question that I have, as they now say, to address, and that is the problem of authenticity; the one word in the title of my talk that is in some respects the most controversial of all. I, I suppose I did agree to the use of the word authenticity in the title, and I believe that the Survey in a very real sense is a collection of authentic language, but what do I, or you, or anybody, mean by authentic language?

Well, in Britain in any case, the question of authenticity in language is now the subject of intense debate. And as far as I can see, the debate is conducted around two questions. First of all, in pedagogical circles the phrase authentic English is typically used in con-

trast with such phrases as simplified English, or textbook English. And, of course, the language, as you'll, as we'll see, of, that is collected in the Survey of English Usage is not simplified in the way that examples are often simplified when writing textbooks to teach a foreign language to people. So in that sense the language we study is authentic rather than textbook language. But, authentic English – or, indeed, authentic language – authentic English can also be contrasted with the idealised English which some theoretical linguists consider to be the object they seek to describe and explain. So we have a contrast between authentic English on the one hand, and idealised English on the other. Language thus idealised is standardised, that is, dialects are not considered. It's the standard language that is investigated. It is regularised in the sense that hesitations, false starts, and mix-ups are not considered, and it is de-contextualised in the sense that the sentence, rather than the text or discourse, is taken as the upper limit of grammatical description.

Now, that is idealised English which is, let's see if I can do this, if I've got, yes [starts writing on the blackboard], which is standardised, regularised and de-contextualised. Okay.

Now, the Survey of English Usage deliberately set out to capture language that was in some important respects not idealised in the way I've been describing, so that the ideas of theoretical grammarians and pedagogical grammarians, too, could be compared with the reality of language use. The language in our files is presented in context, so idealised language is de-contextualised, but the Survey language is contextualised. The basic units of the Survey corpus are texts of about 5,000 running words each, and each grammatical, lexical, intonational or punctuational feature we record is displayed on a slip which can provide 12 to 17 lines of context which is 90 to 120 words of context. That's a considerable amount of context, much more, I might add, than you get in an ordinary KWIC, K-W-I-C, concordance programme on the computer.

So the Survey's language is con-, is contextualised. And all the pauses and slips of tongue and pen have been scrupulously preserved, providing important evidence of the causes and consequences of the great many kinds of non-fluency, so the Survey of English Usage corpus is not regularised. It presents the language, to quote a famous phrase, "warts and all". The well-formed constructions are preserved, but so are the ill-formed, badly-formed constructions, which are very important for psycholinguistic research among other things.

Nevertheless, although the Survey corpus is not regularised and not de-contextualised, it is standardised. It is standardised in several important ways. First, it is a survey of British English, rather than of American English, Australian English, Indian English, etcetera. Second, it is a survey of educated British English, so that while speakers of British English dialects are not excluded as such, and accents other than so called R-P or Received Pronunciation, standard Southern British English pronunciation, are represented, the language collected is intended to be maximally acceptable throughout Britain, rather than restricted to a particular region. And the way we do this is by imposing not a, an explicitly, an explicit dialect standard: "only speakers of some sort of standard dialect", but by imposing a, an educational criterion: only speakers with university education or the equivalent are admitted as valid sources for our texts. That phrase "or equivalent", of course, allows us to take people like the Queen [laughter], who may or may not have the university education, and influential personalities who are admitted as opinion-makers, even though we don't know their educational background.

Third, the Survey of English Usage is a survey of adult English, rather than of language used by or to children; though I personally would like it to include language used by adults to children, it does not. Nor does it use, include the language of children.

And fourth, it is a survey of the English of native speakers, who-, whoever they are, rather than of learners or of those who use En-

glish as their second language. So, this element of idealisation chiefly in the area of imposing standardisation on our sources of text, this element of idealisation not only allowed the Survey's work to be confined within manageable limits, but also led to the assembly of a coherent body of data which could serve as a model for the study of other types of English: American English, Indian English, and so on, and with which these other types of English could then be compared.

The main thing is this, main point is this: the Survey deliberately restricted the range of language users that would provide us with the texts, in order, however, to expand greatly the range of uses of language that we could investigate. So if the Survey corpus is deliberately bounded with respect to the users of English represented, it is deliberately diverse with respect to the uses of English recorded, which are classified according to many of the factors basic to any ethnography of communication.

And at this point, it might be a good thing to have a look at our text tree, to have a look at some of the texts that we actually take. There are two versions of this tree that I have, a simplified one and a fuller, more complex one.

I have only ten copies of this simplified text tree, perhaps you could, you, take one or two per Z, yes. Thank you. So several of you can look at one copy.

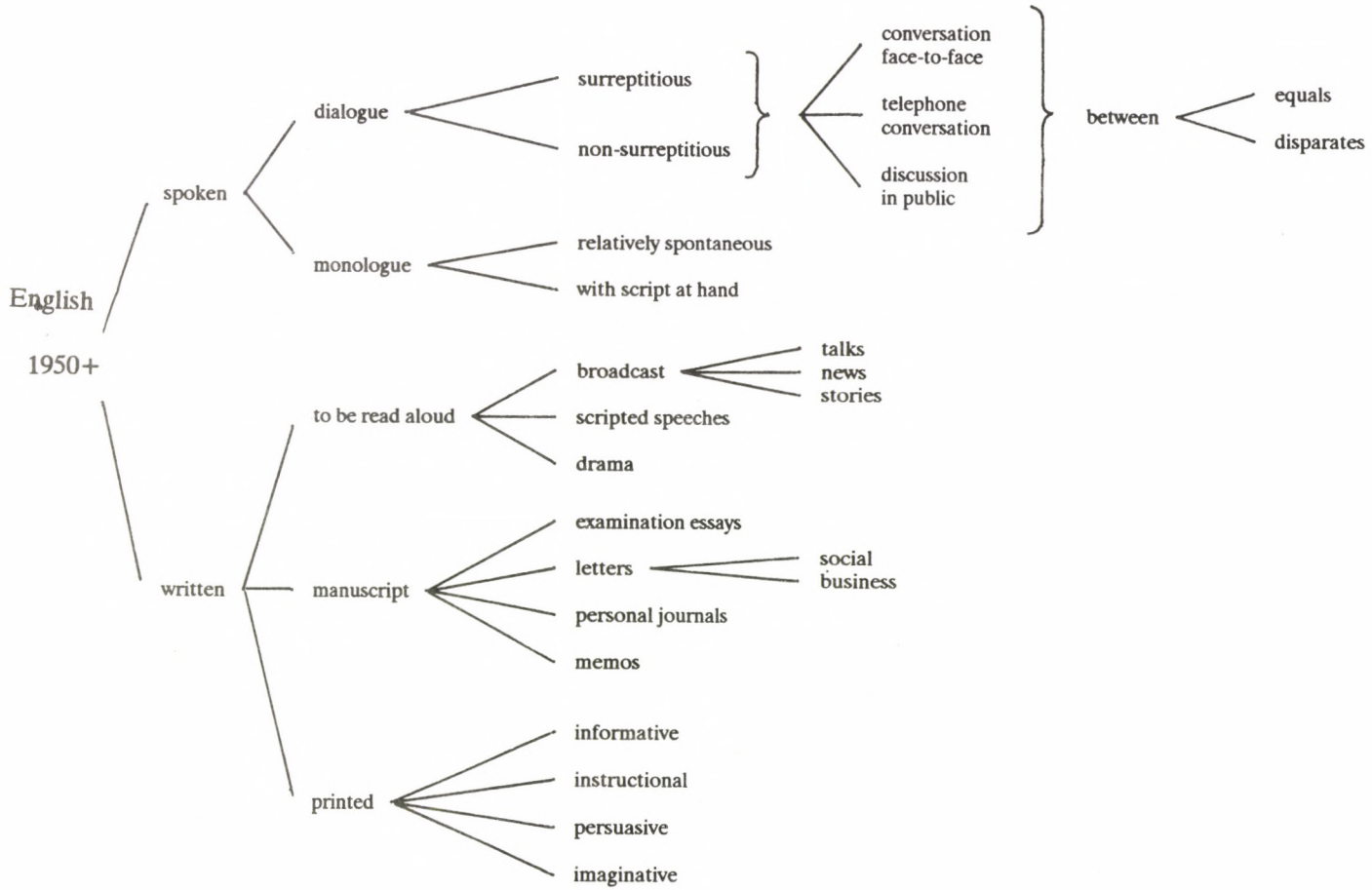
So here is our first version of the Survey text tree,[1] a simplified version, illustrating how, though we have restricted our potential users of language who are the sources of our texts, we have, nevertheless, tried to collect a wide variety of uses of the language. As you can see, the Survey is pretty synchronic, since none of our texts is allowed to date from before the 1950s, it is, therefore, contemporary English that we collect.

---

[1] Table 1 (see page 6)

Table 1

Corpus of the Survey of English Usage



The basic division of our corpus is between spoken language and written language. The spoken language is, in turn, divided into dialogue and monologue, the dialogues interestingly into surreptitious and non-surreptitious; that is, dialogues whose participants did not know they were being recorded at the time, though their permission was usually asked afterwards, and those whose participants did know they were being recorded at the time, and these dialogues have been divided into face-to-face ones, face-to-face conversations, dialogues at a distance over the medium of the telephone, and public discussions, such as panel broadcasts on the BBC, which, where, of course, the participants knew they were not only being recorded but that they were broadcasting to a large public. And, furthermore, these dialogues are divided on the basis of participant relations. They're a very difficult thing to do, but roughly they are divided into dialogues between equals and dialogues between disparates, where disparity, in the best sociolinguistic fashion, is a function of such factors as differences of age and differences of social status.

Then, we have spoken texts which are monologues and these, in turn, are divided into those that are relatively spontaneous and have almost no preparation in advance, and those where there is some sort of script or written notes at hand.

So much for the spoken texts, of which there are meant to be one hundred 5,000-word texts, a total of 500,000 running words.

The other half of the Survey, the other 500,000 words is a hundred written texts. And these are divided into three categories: texts that are written to be read aloud, including broadcasts and scripted speeches and drama; manuscript texts of various kinds, including letters, diaries, exam scripts written under the pressure of time. (And it's very interesting: I believe that the Survey of English Usage is the only corpus in the whole world of any language which has deliberately included written material that is not printed, that is handwritten or typed, and the most important thing about this material is that it has not been subjected to any-, to editing by anybody

else, by a subeditor, a publisher or anything else, so we can get material straight from the author's pen or typewriter.) And then, of course, we have your standard, and in some respects rather boring, but interesting, staple, the printed texts, which consist of a variety of excerpts from novels, excerpts from newspapers, excerpts from legal documents of various kinds, nonfiction books, and the like. And these are divided in various ways, including those that are primarily informative or expository, those that are instructional,[2] some journalism, and the imaginative prose of novelists. Now you'll see that the criteria for dividing the corpus in this way is not the same, the criteria are not the same throughout. Thus, for example, we do not on the whole divide the spoken texts up into [=by] their communicative intention. We do not have categories of instructional spoken English, persuasive spoken English, mandatory, org-, ordering spoken English in the way that we divide up the printed texts. Nor, for example, do we or can we make a distinction in our printed texts according to the participant relations of writer and reader. We do not attempt to divide newspaper articles, or articles in learned journals or novels up into those where the writer is on intimate terms with the reader, and those where the writer is not. I don't see how really we could make that distinction here, even though we tried to make the distinction up here [refers to Table 1] in the spoken texts.

This is all very simple and straightforward, is it not? And yet, we have found, even when assembling our corpus of material, that no distinction in language is ever simply a matter of black and white, or ei-, either-or. Thus, for example, you will note that among our spoken texts we have those that our, th-, are delivered with script in hand, and among our so called written texts we have those which are written to be read aloud. And you might well ask what the difference is between those two texts. In fact, that is a question that caused me a great deal of trouble when I looked in more detail at

---

[2] also those intended to persuade (lecturer's written addition)

the Survey texts. And I have here once again ten copies, oh, perhaps even more, of an expanded version, a detailed version, of our Survey text tree,[3] which will enable you to look at the specific texts that we are collecting.

Table 2

List of Texts Included in the Survey of English Usage

WRITTEN TEXTS:

Text Number	Category	Description	Date
	WRITTEN FOR SPOKEN DELIVERY		
W.1.1	Talks	BBC 3rd Programme talks on art	1964, 65
2	Talks	BBC talks on science subjects	1965, 71
3	Talks	BBC talks: autobiographical reminiscences	1964, 72
4	Talks	Exhibition guides	1984
5	Talks	BBC talks on aspects of teaching	1965, 72
6	Talks	BBC Week's Good Cause/ Thought for the Day	1984
W.2.1	News	BBC Home Service Radio	1964
2	News	BBC Market Trends and Market Reports	1964, 70
3	News	Radio 4 News and "The World at One"	1971, 72
4	News	ITV News	1984
W.3.1	Stories	BBC "Book at Bedtime," recorded	1964, 65
W.4.1	Formal scripted speeches	Speech from the Throne and Prime Ministers' Speeches	1970, 71
2	"	Public Orator, Foundation Day	1968, 71
W.5.1	Drama	"The Hotel in Amsterdam"	1968
2	Drama	Radio 4 "Afternoon Theatre"	1974
3	Drama	Radio 4 "Midweek Theatre"	1974
4	Drama	BBC TV plays: "Terra Nova" "Leaving"	1984

[3] Table 2 (see pages 9-14)

Text Number	Category	Description	Date
	NON-PRINTED MATERIAL		
W.6.1	Continuous writing	Examination essays, English literature	1963, 65
2	"	Examination essays, English literature	1965
3	"	Examination essays, geography	1973
4	"	minutes of meetings	1983-5
5	"	pressure group newsletters	1985
6.6	see W.13.5		
W.7.1	Letters – social, intimate	mother to daughter	1962-5
2	"	student (male) to family	1962
3	"	letters to friends	1962-5
4	"	student to girlfriend	1963
5	"	letters to friends (female)	
6	Letters – business	typed referees' letters	1976
7	"	typed letters of application	1977
8	"	handwritten letters of application	1976-7
9	"	bank-manager to client	1959-65
10	"	solicitor to client	1963-5
11	"	solicitor to counsel	1976-7
12	"	medical correspondence – consultants	1978
13	"	medical correspondence – to GPs	1978
14	"	medical correspondence	1978
15	"	handwritten business letters	1975-7
16	Letters – printed/ mimeoed	mass distribution letters	1984
17	Letters – pre-publication	To the Editor, *London Rev. of Books*	1986
W.7.31	Letters – social, intimate	letters between undergraduates (female)	1976-8
32	"	letters between undergraduates (female)	1975-7
W.17.1	Letters – business	typed referees' letters	1977
2		handwritten referees' letters	1977
W.8.1	Journals	February – March	1966
2	"	January – May	1965
3	"	March 1951 & March 1964	
	PRINTED MATERIAL		
W.9.1	Learned arts	P. F. Strawson, *Individuals*	1959
2	"	G. W. Bromily, *Essay in Cristology*	1956
3	"	G. Kitson Clark, *The Critical Historian*	1967
4	"	W. Nowottny, *The Language Poets Use*	1962
5	Learned sciences	E. C. Barnett, *Climatology from Satellites*	1974
6	"	T. P. Bayliss-Smith, *The Ecology of Agricultural Systems*	1982

Text Number	Category	Description	Date
7	Learned sciences	J. Z. Young, *The Life of Vertebrates*	1950
8	"	G. E. Bacon, *Neutron Diffraction*	1955
9	"	H. N. Y. Temperley, *Changes of State*	1956
10	"	G. H. Williams, *Homolytic Aromatic Substitution*	1960
11	"	C. A. E. Goodhart, *Monetary Theory and Practice*	1984
W.10.1	Instructional writing	G. Stanton, *Handyman's Handbook*	1966
2	"	G. Dawson, *Tackle Sailing This Way*	1959
3	"	F. Fairbrother, *Roses*	1958
4	"	miscellaneous instruction manuals for various home appliances	1973, 78
5	"	instruction manual for the SCIENTEX word-processing system	1982
6	"	Inland Revenue instructions	1985
W.11.1	General non-fiction	R. Hoggart, *The Uses of Literacy*	1957
2	"	*New Statesman*, essays	1966, 68
3	"	M. P. Lockley, *Wales*	1966
4	"	feature articles, *Guardian* and *Times*	1972
5	"	D. M. Hill, *Participating in Local Affairs*	1970
6	"	M. Sullivan, *Chinese Art*	1973
7	"	G. R. Taylor, *The Great Evolution Mystery*	1983
8	"	articles from specialist magazines	1984
W.12.1	Press – general news	*Times*	1964
2	"	*Daily Express*	1964
3	"	*Guardian*	1968
4	"	*Daily Telegraph*	1968
5	Press – specific news	*Guardian* football reports	1968
6	"	*Daily Telegraph & Financial Times* financial reports	1968
W.12.7	Local press – general news	miscellaneous	1984
8	Press – editorials	miscellaneous	1984
W.13.1	Administrative and official language	University of London Reorganisation 1964 – 1966	1966
2	"	Company reports, *Times*	1968
3	"	leaflets and pamphlets	1980-3
4	"	printed notices	1985-6
5	"	handwritten notices	1987 renamed W.6.6
W.14.1	Legal and statutory lg.	Health Services and Public Health Act	1968
2	"	Law Commission Report	1983
3	"	leases	1980-1

Text Number	Category	Description	Date
W.15.1	Persuasive writing	M. Stockwood, *Bishop's Journal*	1964
2	"	sermons	1957, 60, 65
3	"	party manifestoes	1964, 66
4	"	advertisement features, *Sunday Times* and *Evening Standard*	1972
5	"	junk mail	1983–
W.16.1	Prose fiction	R. Lehmann, *The Echoing Grove*	1953
2	"	M. Bradbury, *Eating People is Wrong*	1959
3	"	A. Waugh, *The Foxglove Saga*	1960
4	"	D. Beaty, *The Proving Flight*	1956
5	"	L. P. Hartley, *A Perfect Woman*	1955
6	"	A. Wilson, *Anglo-Saxon Attitudes*	1956
7	"	L. Davidson, *The Rose of Tibet*	1962
8	"	S. Hill, *Strange Meeting*	1971

SPOKEN TEXTS:

(f = female, m = male, NS = non-surreptitious recording, part. = participants)

Text number	Category	Number of part.:	Additional information on participants etc.:	Date recorded:
S.1.1	surreptitiously recorded conversation – intimates	2	m	1964
2	" "	2	m (3 conversations)	1963-5
3	" "	3	2f, 1m (2 NS)	1965
4	" "	2	m	1969
5	" "	4	f	1967
6	" – intimates/equals	2	m, f	1964
7	" – intimates	3	m (1 NS)	1972
8	" "	3	f	1969
9	" – intimates/equals	4	3m, 1f (1 NS)	pre 1966
10	" "	3	2f, 1m (2 NS)	1975
11	" "	3	2f, 1m (2 NS)	1975
12	" – equals	4	2f, 2m (2 NS)	1975
13	" "	3	2m, 1f (1 NS)	1975
14	" – intimates/equals	3	2m, 1f (2 NS)	1976
S.2.1.	" – intimates	2	m (3 conversations)	1953, 64
2	" – equals	2	m (2 conversations)	1969
3	" "	3	2m, 1f (1 NS)	1974
4	" – intimates/equals	4	3m, 1f (2 NS)	1970 (?)
5	" "	3	2m, 1f (1 NS)	1974

Text number	Category	Number of part.:	Additional information on participants etc.:	Date recorded:
S.2.6	surreptitiously rec. conversation – intimates/equals	4	m (1 NS)	1974
7	" "	3	2f, 1m (1 NS)	1975
8	" – intimates	4/2	3m, 1f (2 NS / 1 NS)	1975
9	" – equals	3	2m, 1f (2 NS)	1974
10	" "	4	2m, 2f (2 NS)	1975
11	" – intimates/equals	2/4	2m, 2f (1 NS / 2 NS)	1975
12	" "	2	f (1 NS)	1975
13	" "	4	2m, 2f (2 NS)	1976
14	" "	3	2f, 1m (1 NS)	1976
S.3.1	" – disparates	3	2m, 1f (3 conversations)	1961
2	" "	2	m, f,; 2m (3 conversations)	1973-5
3	" "			1971 (?)
4	" "			1971 (?)
5	" "	3	m (1 NS; 2 conversations)	1961
6	" "	5	m (1 NS)	1974
7	" "	3	2m, 1f (1 NS)	1984
S.4.1	non-surreptitiously recorded conversation – intimates	2	m, f	1969
2	" "	2	m, f	1971
3	" – equals	4	2f, 2m	1972
4	" "	4	3m, 1f	1975
5	" – intimates	3	2f, 1m	1976
6	" – equals	4/5	2m, 2f / 3f, 2m (2 conv.)	1976
7	" – intimates	3	2f, 1m	1976
S.5.1	non-surreptitiously recorded conversation – equals	5	BBC "Any Questions"	1959
2	" "	5	BBC "Brains' Trust"	1958
3	" "	3	BBC "What's the Idea "	1961
4	" "	5	BBC "Any Questions"	1958
5	" "	5	BBC "Any Questions"	1960
6	" "	4	BBC "What's the Idea"	1961
7	" "	4	BBC "A Word in Edgeways"	1970
8	" "	2	m, f	1971
9	" "	2	m, f	1971
10	" "	2	m	1971
11	" "	2	m	1976
12	" – intimates/equals		Choir Committee meeting	1985
13	" "		Academic Council meeting	1986
S.6.1	" – disparates	2	BBC interviews (3 conv.)	1966
2	" "	2	m, f	1961
3	" "	2	BBC interview	1974
4	" "	3/2	2m, 1f/2f (part BBC)	1973, 75
5	" "	4	BBC interviews	1975
6	" "	1	BBC recording	1974
7	" "	2	BBC interview	1971
8	" "	4	Psychiatrists' discussion grp.	1977
9	" "		computer use instructions	1985, 87

Text number	Category	Number of part.:	Additional information on participants etc.:	Date recorded:
S.7.1	telephone conversations – intimates			1961, 67
2	"	"		1975
3	"	"		1975
S.8.1	"	– equals		1975
2	"	"		1975
3	"	"		1975
4	"	"		1975 (?)
S.9.1	"	– disparates		1975-6
2	"	"		1975
3	"	"	Answerphone	1975
4	radio & phone conv. – disparates		phone-in on investments	1985
5	dictated letters			
S.10.1	spontaneous commentary	– sport 4	cricket	1964
2	"	" 2	football	1971
3	"	" 2	boxing	1960
4	"	" 2	horse-racing (TV)	1960
5	"	– other 5	Churchill's funeral	1965
6	"	" 2/3	royal wedding	1973
7	"	" 1	launching of ship; royal visit; physics demonstration	1960-76
8	spontaneous commentary	– other 1	BBC "Living World;" physics demonstration	1976
9	"	" 1	physics demonstration; biology demonstration	1976
10	"	– sport	BBC TV tennis at Wimbledon	1984
11	"	– other	TV cookery demonstration	1986
S.11.1	spontaneous oration	3	legal cross-examination	1967
2	"		speech, Fellows' dinner	1974
3	"		BBC "My Word"	1961-2 (?)
4	"		House of Commons question time	1975
5	"		House of Commons debate	1975
6	"		House of Lords debate	1986
S.12.1	prepared oration		sermons	1965
2	"		lectures	1965, 67
3	"		address to court	1966
4	"		judgments	1966 (?)
5	"		speech, party conference	1972
6	"		lecture	1972
7	"		UCL Foundation Oration	1973

I don't propose to examine this detailed tree in detail for lack of time, but when my few copies have been distributed, I hope to call your attention to at least one of the kinds of problem, one or two of the kinds of problem that this sort of classification or taxonomy of texts present[s].

I wonder, yes. May I possibly, just for the moment, trade this [his handout] for that [somebody's handout in the audience] because unfortunately this has an extra sheet. Those of you who have a three-page version of this document, as, I think, most of you have, might like to look at the very last page, and the category labelled *S.12: prepared orations*. We have, as you see, seven prepared orations, each of 5,000 words: some sermons, lectures, addresses to the court, judgments and, and so on. These, as you saw in the earlier handout, are supposed to be texts backed up by, supported by, a certain amount of writing in the form of written notes. But by contrast, on the first page, you might turn your attention to category *W.4: formal scripted speeches*; a Speech from the Throne opening Parliament is one of them, the other is the Public Orator of University College London, presenting degrees, honorary degrees, to distinguished people.

So, if you compare the S.12 texts with the W.4 texts, ((somewhere)) you get almost a minimal pair. Text S.12.7 is the University College London foundation oration. Text W.4.2 is the Public Orator of University College awarding honorary doctorates. One of those texts is regarded as a text that was created in writing, and realised in speech. That is text W.4.2. The other is regarded as a text that was both created and realised in speech. That's text S.12.7. Even though it had notes. And yet, if you look at the headings of the texts, they're very very similar. And when I saw that, I was ast-, I was very worried. On what basis can that distinction be made?

Well, the W.4.2 text was speeches by Joel Hurstfield, the, formerly the Public Orator of University College. The S.12.7 text was a speech delivered by Jonathan Miller, a name that is one to conjure

with in Britain,[4] though perhaps not here. Fortunately, I had been involved in both texts. I transcribed Joel Hurstfield's text, and I checked the transcription of Jonathan Miller's text. And they were different. It was very clear to me that Joel Hurstfield was, in fact, reading from a thoroughly written-out text. He read it very well. He made it sound lively. But it was, nevertheless, completely written out as a text before he delivered it. But as for Jonathan Miller, it was equally clear that although he had a text, he departed from it in many ways. He told stories, he told jokes, he changed his presentation very much in response to his public. He is, in, among other things, a man of the theatre. So it seemed to me, after a good deal of anxiety, that this class-, this distinction and classification was justified. And it shows very well that even so basic a distinction as the difference between S texts, that is, texts with their origin in speech, and W texts, that is, texts with their origin in writing, even so apparently straightforward a distinction, is not so straightforward after all. And here we have almost a minimal pair showing how fine the distinction can be between an S text and a W text.

Of course, the obvious question, which I am sure you're all thinking about now, is what kind of text I am generating as I speak to you. Have you any ideas about what sort of text I am generating, and how it might be classified in the Survey of English Usage? [pause for reply; some laughter]

Well, it would probably be an S.12 text, rather like Jonathan Miller's. It is true that I have some written backup, but as you can see, what I have is notes, rather than a completely written out text. And that means that most of what I'm doing is S.12: a formal, script-, a, a, a speech backed up by notes, but nevertheless created as a text in the act of talking to you.

---

[4] Dr Jonathan Miller, CBE, MB BCh, DLitt: polymath and medical doctor; television producer and director of plays and operas.

But there are some problems. A few minutes ago I was reading out a, some excerpts from an article that I wrote a few years ago about the Survey of English Usage. So what was I doing then?

It seems to me that at that point I was performing some sort of W.4, let us say Z, some, some ki-, some one of these texts W.1 to W.5. Then I was realising in speech a text that very very clearly had its origin in writing. So in the theatrical performance that I am now giving to you you can see some of the problems of text classification when you are serious about it. Because one of the things that we found over the years at the Survey is that texts can be of mixed type. I am producing a text that is mostly an S.12, but that has elements in it, from time to time, of W.4. So how can such a text be classified? In f-, point of fact, there are a number of Survey texts which have this mixed character, and whose classification depends on what they are mostly. So this is mostly an S.12 text, with some W.4 elements in it.[5] And that illustrates a numb-, some of the problems simply at the level of the text taxonomy.

And there are other problems that have emerged. Over the twenty-, the. First of all, why, why has the Survey taken so long? We've been going for twenty-five and more, twenty-seven and more years, and we still haven't collected all our texts. Well, that's because some of the texts are very difficult to collect, there are many problems, for example, in collecting telephone texts, which I won't go into now. And also because of a deliberate decision that we took a long time ago, which was that the tax-, the task of collection and the task of analysis should proceed simultaneously. That means that from the early 60s we had at least a few texts that we had analysed completely, and that were available for researchers to use, and

---

[5] That is true of the text that people heard in 1988, its tape-recording, and the transcription of the recording made by Andrea Reményi. But the version of that text published here will have been edited, however slightly, thus introducing an element of W.9, which is probably where the Survey would classify its published form. (lecturer's written note)

scholars to write papers about. So instead of collecting all our texts at once and then analysing them, we always had a little bit of analysed material, a growing amount, in fact, that people could use.

Now, because of this historical dimension to the Survey's activities we have seen some interesting changes over the years. Thus, for example, some, some kinds of language use that we thought we were going to, were very important, have proved not to be. Originally we were going to collect lots of, of diaries, people's diaries, as a kind of manuscript, or non-printed text. We found that very few people keep diaries any more, in contrast to the practice of times past, where diaries were a very important genre. On the other hand, we have seen the emergence of new types of language use: the phone-in programme broadcast, where somebody stands in front of a microphone in a radio station, and people telephone to him to ask questions, make comments and so on, on one issue or on a variety of issues. That is now a very prominent feature of broadcasting in the United Kingdom. It probably was not so common in 1959. We have seen, above all, the rise of television. And one of the things that we're doing now in our last bit of text collection is to, to parallel a lo-, a, a, a lot of our early radio texts with similar texts recorded from television. What's the difference between a radio news broadcast and a television news broadcast, etcetera? Between a radio play and a television play? Now, there are problems, too, in all this, including the, the way in which one type of text fades or changes into another type of text. Text boundaries are not necessarily clear-cut, as we saw in the distinction between S.12 texts and W.4 texts.

There are other problems, too, including cross-classification and mixed classification and what classification. We've seen the problem of mixed classification in the text that I am generating in speaking to you: an S.12 text with W.4 features. And I assure you that there are many such problems in that respect. For example, in news broadcasts. The news broadcasts were originally classified simply as texts that were written to be spoken. That, I think if you look, you'll find

them in te-, in, as W.2s, or something. Which means that when the Survey was organised the typical news broadcast was fully scripted, and according to old BBC hands, a man, because in those days it was only men, came into Broadcasting House, dressed in a dinner jacket with a black tie, and sat down in front of a microphone, and read the news. In fact, he was called, and to some extent he's still called, a newsreader. And he would say: "This is Al-, this is the BBC, this is Alvar Liddell, and here is the six o'clock news". And you got the six o'clock news. That's why our news broadcasts are W.2s.

But now, when we take, say, television news broadcasts, they're completely different. Not only, I think the, the presentation by the so called anchor people is still probably fully scripted, but there are also lots of outside broadcasts, there are interviews with victims and executioners, so to speak. People, the-, there are reports from, from correspondents, as the bullets whistle round their heads. And those are clearly not scripted. Yet, despite the fact that news broadcasts are now very different from what they used to be, we still have to take account of them, record them, and classify them somewhere in our text tree.

And then there is the problem of some texts that's, that are extremely difficult to classify at all. What do we do, for example, with dictation? Somebody dictates something to a secretary, who then types it. What, where do we put that in our text tree? Well, if you look at the text tree, you'll find that it's extremely difficult to classify. Is it a kind of conversation? Is it a ki-, conversation between the boss and the secretary? Is it a kind, is it a sort of monologue? Well, if it's a monologue, there is, but there is somebody, if it, if it is a kind of monologue, what kind of monologue is it? The most important oddity about dictation is that there is a third person involved in the communicative activity, but that person is not present. That person is the recipient of the eventual printed letter. So, just as in the early days we realised that there were texts that were written to be spoken, for example, a fully scripted speech, so we now realise

that there are texts that are spoken to be written, namely texts that are dictated. But it's extr-, I frankly do not know, I don't think that there is any convenient place in the entire Survey t-, text tree where ordinary dictation can conveniently be, be handled.

And then, as I said a, a few seconds ago, there is also the problem of cross-classification. If you look at the text tree, the full text tree, you will see that w-, that there are, there is a category, I think, called S.10, which is sup-, which is officially labelled 'spontaneous commentary'. All right?

So, very easy: you comment on, or, as they now say, commentate, on what is happening in front of your eyes. So, you give a commentary on a boxing match, or you give a commentary on a boat race, or you give a, a, a commentary on a royal occasion, like one of the innumerable marriages of princesses to princes that go on all the time in Britain, and it's happening in front of you, so clearly, we thought, if you're talking about what's happening in front of you, it's spontaneous, it can't be written out. Little did we suspect that it wasn't so simple as that. And that became, that was forcibly borne home to us when we finally got round to recording a cooking broadcast on television. This is a very popular genre in Britain, and may well be here, I don't know. There is somebody who stands in front of the camera, and says: "I am now going to make an omelette. So I break an egg, I put it into this bowl, I put some milk into the bowl, I put some spices into the bowl, and mix it all up together, I put it in a pan, I cook it." Now, that's commentary. The, the speaker is talking about what is happening. But it's commentary which is about an activity that is directed by the speaker herself or himself. And, in the case of these cookery broadcasts, there is every reason to suppose that the presenter has done it all before [loud laughter]. In that case the question arises, is the commentary spontaneous? One can go further: is the, the cookery person, the cookery expert actually spouting even rehearsed lines from his or her own past experience, or are they, shifting now to the plural, reading

their commentary off the so called auto-cue, which appears just above their head, and which is known popularly in Britain as the idiot-board [laughter]. It is said that certain presidents of the United States [laughter] have made ((extensive)) use of such devices. That need not appear in the printed version. So, we even here find that in a mixed classification, like spontaneous commentary, we find that there can be commentary, definite commentary which is not necessarily spontaneous. And that creates another problem of, of classification.

Looking back on the history of our enterprise, we find that a, a combination of primitive technology in 1959 (when we had very heavy tape-recorders, and they had to be plugged in), the problem of getting texts whose speakers did not know they were being recorded, but would then grant us permission to use their texts, a desire to know the background of our speakers, and so on, imposed constraints on the types of text we could collect. There is relatively little passion in our texts. We have few arguments, we have few love dialogues in speech, although we do have a collection of love-letters, which somebody donated to us. There's very little of the language of work, the language of the assembly line or the, the board meeting. There is very little in the way of outside recording, in, in the street, as it were. We're trying now, using more modern means of production and recording: cassette recorders, and so on, to make up for that. And we are getting a wider variety of texts within our classification as the Survey nears completion.

We now co-, and g-, so much for assembling the corpus. What about analysing the corpus? Well, when we have assembled the texts, a-, and transcribed the spoken ones, and so on, we then collect the prosodic features, the punctuation features, the closed-category linguistic items, and some essential grammatical structures. So, for example, we have a method of transcribing spoken texts, in which fundamental factors of intonation, rises and falls, changes in loudness and speed, and so on, are transcribed and recorded. We collect

for the written texts what corresponds to prosody in the spoken texts, namely, punctuation: full stops, quotation marks, exclamation marks and commas and hyphens. For all texts we collect closed-category lexical items, things like *the* and *a*, prepositions like *in* and *at* and, and *on*, and so on, modal verbs: *can*, *may*, *must*, *should*, and many phrases: *in front of*, *due to*, *because of*, and then we collect a variety of grammatical structures, as well.

And, what I would like now to distribute to you is, and I have a f-, more copies of this than of the preceding handouts, [it] remains only to find them, yes, the Survey text tree, this comes in two pages like this, and in some cases, the two sheets are, distribute, are together, and in other cases they are separate.

Now, this [the distributed handout][6] just shows the grammatical or quasi-grammatical structures that the Survey collects. It does not show the punctuation marks, the, it doesn't have a list of all the closed-category lexical items that we collect.

Table 3

Grammatical Categories of the Survey of English Usage

1. Exclamatory Noises
2. Formulas
3. Abbreviated Forms
4. Cardinal Numbers
5. Ordinal Numbers
6. Zero Articles before Numbers
7. Zero Articles before Nominal Groups

[6] Table 3 (see pages 22-24)

8. Nominal Groups
9. Apposition I
10. Apposition II
11. Names
12. Vocatives
13. Adjective Sequences
14. Comparatives and Superlatives
15. Adverbs
16. Negation
17. Finite Verb Groups: Simple
18. Finite Verb Groups: Complex
19. V + *ing*
20. V + *to*-infinitive
21. V + *to*-less infinitive
22. V + *ed* (Active)
23. V + *ed* (Passive)
24. V + Obj + *ing*
25. V + Obj + *to*-infinitive
26. V + Obj + *to*-less infinitive
27. V + Obj + *ed*

} Complex Finite Verb Groups

28. Nonfinite Verb Groups: Simple – *ing*
29. Nonfinite Verb Groups: Simple – *infinitive*
30. Nonfinite Verb Groups: Simple – *ed*
31. Nonfinite Verb Groups: Complex
32. V + *ing*
33. V + *to*-infinitive
34. V + *to*-less infinitive
35. V + *ed*
36. V + Obj + *ing*
37. V + Obj + *to*-infinitive
38. V + Obj + *to*-less infinitive
39. V + Obj + *ed*

} Complex Nonfinite Verb Groups

40. Verbs + Adverb/Preposition
41. Words + Adverb/Preposition
42. Prepositional Phrases

43. Preposition + Noun + Preposition I
44. Preposition + Noun + Preposition II
45. Correlatives
46. Complex Prepositions
47. Finite Verb Constructions
48. Nonfinite Verb Constructions: *– ing*
49. Nonfinite Verb Constructions: *– infinitive*
50. Nonfinite Verb Constructions: *– ed*
51. Verbless Constructions
52. Ellipsis
53. Anacolutha
54. Grammatical Obscurities
55. Exponents of Subject
56. Exponents of Complement
57. Exponents of Multiple Complement
58. Anaphora
59. Zero Subordinators
60. Reference or Concord
61. Direct Speech
62. Indirect Speech
63. Erlebte Rede
64. Extra-Idiolect
65. Foreign Phrases

You will see that what we aim at is a theory-neutral classification. So that, whether you are a, an adherent of Chomsky or of Fillmore or of Coseriu or of any other, or indee-, o-, of, of any other linguistic theoretician, you can come and use the files of the Survey, and find material of use.

And so we collect very traditional things, like nominal groups and vocatives, and verb phrases, and, what is traditional in English, but apparently is not quite at the centre of attention in Hungarian, prepositional phrases. And you can look at the list at your leisure.

Now, we have tried to make this list theory-neutral, so anybody can use it regardless of their theory perspect-, their theoretical perspective. But it is clearly not, or is it? It is cle-, clearly not language-neutral. The, the categories we have collected seem to be those that are of particular relevance to the grammatical analysis of English. And one of the things that I hope you'll tell me in a little while is whether any of these categories would be relevant to, for example, a corpus concordance of Hungarian. Some of them might be, some of them might not be, others might well have to be added. One thing that y-, you can see here is that the Survey, greatly to my regret as a lexicographer, has not traditionally collected very much morphology, either inflectional or derivational. We may be able to make up for that lack, now that we are embarking on a programme of full-scale computerisation. But we have very little of it in our basic, we have very little morphology in our basic files.

However, some general problems emerge from considering this, a list of categories. First of all, the categories have to try to achieve the right delicacy, the right level of specificity. They cannot be too specific, they cannot be too general. Thus, for example, those of you who're interested in English may know that a very important thing in English is the so-called phrasal or prepositional verb, things like *give up*, *run across*, *look down on*, *put up with*, etcetera. You will notice that the Survey does not collect phrasal verbs. What it does do is collect, somewhere around here you can see it, there is a category called 'verb plus adverb or preposition', which, strangely enough, I cannot find, though it is here. It's 40. 'Verb plus adverb/preposition', thank you very much. Now, that looks very much like phrasal verbs, doesn't it? But it's not intended to be. It is intended to cover a whole range of material, ranging from the purely idiomatic combinations (the *give up*: *I, I have to give up smoking*; *put up with*: *I won't put up with their behaviour any more*, and so on, the purely idiomatic combinations) to p-, perfectly simple and straightforward combinations, like *I looked down the telescope* (as well as *I*

looked down on my inferiors); *I walked down the stree-*; *I looked*; *I, I ran across the road*, perfectly transparent (as well as *I ran across a friend yesterday by chance*, which is an idiomatic combination). In that way, the category 'verb plus adverb/preposition' allows us to look at the whole range of phenomena from the hard core, idiomatic combinations, to the almost literal combinations. And in that way we can come up with some new insights into the nature of the phrasal-verb construction in English, and what distinguishes them, if anything, from straightforward combinations of verb and particle.

I, as, in the absence, since I haven't got all day, I will not go into more details about how we have to achieve the right level of analysis.

Nevertheless, we have found over the years that there is a fruitful tension between the theory, the theoretical framework with which we started, however minimal that was, and the data that we have accumulated over the years. Thus, for example, if you look at our list of closed-category items, which I have not included in the handout, but of which there are 408, you will see that some of them are phrases of the form *in front of*, or *due to*; groups of words that function, like *due to* or *because of* or *in front of*, groups of words that function as prepositions. They function very much like single word prepositions in English. Originally, each one of those groups of words had a file of its own. There's a file for *due to*, there's a file for *because of*, there's a file for *in front of*. But we gradually came to realise that this kind of construction in English is very productive. We have *in keeping with*; in American English *in back of*, as well as *in front of*. We have not only *because of*, but *owing to*, not only *due to*, but the *up to* of *They can earn up to five thousand pounds a month*. And so we then produced several new categories, one of which was 'prep-noun-prep combinations', into which we put things like *in keeping with*, *in view of*, *in light of*, and so on. The other [was] 'complex prepositions', like *owing to* and, and *up to*, and many

others. So you will find in the Survey system of grammatical classification itself evidence of hi-, of historical changes and growing insight based on a fruitful conflict between theory and data. We started out with a few files for particular prep-noun-prep combinations, particular complex prep, preposition combinations. But now we have a more general file into which the whole range of expressions formed from prep-noun-prep, or from, o-, of complex prepositions of various kinds can be placed as well as the original file[s]. That shows how our own theoretical judgments have been influenced by the actual data that we found in our texts.

All right. And, of course, even in the case of, of individual lexical items we keep finding new ones that we want to classify asd [=as] closed category. Wi-, within the past year we came across the first example in the [=a] Survey text of *albeit*, which, I suppose, is some kind of very formal conjunction or adverb or preposition, or God knows what. But it's clearly a closed-category grammar word, rather than a fully lexical item. We had no file for it because we'd never, we did, it'd completely slipped our mind, we'd never come across any examples of it in our texts. But in a legal text we found one, so, of course, we created an, a new file for it.

I now proceed to the use of the text [=corpus], having discussed the assembly of the corpus, and the analysis of the corpus. How is this used?

Well, for that, I think, I can give you a, by way, I think I'll come to this s-, last handout in just a moment.

How can the analysed Survey corpus be used? It can be used as a corpus, it can be used as a a concordance. It can be used to compare registers, or different uses of language. So, for example, we can compare racing commentaries and cricket commentaries. And we find some very interesting differences between quick sports and slow sports. Quick [=continuous-action] sports, as you might, like racing, as you might expect, have a lot of progressive forms in English: *He's going up*, *They're moving up on the outside*, etcetera. Slow

[=repeated-action] sports, like cricket, ((a)) very slow sport ((for)) some, have more simple forms: *He steps to the pitch, raises his arm, He bowls to the ball*[7] by contrast to *One horse is gaining on another.* We can also use the corpus or concordance to look at the language as a whole. Thus, for example, when Professor Frank Palmer was writing his book on the modals in English, he spent one morning at the Survey of English Usage looking at our files of *may* and *can* and *should* and *must* and said that it had changed his life completely [laughter]. I'm sure that Professor Sidney Greenbaum's life was also changed completely by the vast amount of material that the Survey contains on adverbs, and on other realisations of the adverbial in English, because we have a whole file in which all adverbs are taken, and Professor Greenbaum wrote a celebrated book on adverbs and other adverbial constructions in English.

We can use the Survey to compare varieties or dialects of the language, but in order to do that we need parallel corpuses or corpora. There are now plans, for example, to do a spoken corpus for American English, which curiously, for all practical purposes, doesn't exist. I mean there are some historical things, but. So, a-, and whi-, which will be comp-, comparable with a corpus of spoken British English that we will prepare especially for that purpose. There is also a New Zealand corpus in preparation, there are two Indian English corpora in, in preparation, and so on.

And we can use, perhaps we can use the, our corpus to compare languages. Thus, there is a well known Serbo-Croatian–English contrastive project, which uses Survey material and Serbo-Croatian material to compare. I wonder how far a, a Hungarian–English contrastive project, if there's any interest in it, could be organised in the same way. In principle we can use the corpus to analyse language-

---

[7] or: *He runs up to the wicket, He bowls, The batsman plays forward* (lecturer's written addition)

learners' errors, by comparison with what our native-speaker sources produce. But that hasn't been done very much.

In dictionaries the Survey material has been of great use, first of all, to improve the treatment of function words or closed-category items. One of mo-, my most horrible experiences a l-, as a lexicographer working on the *Longman Dictionary of Contemporary English* was to find, when I was able to send a colleague to look at the Survey file of *of*, *o-f*, a, an English preposition, I, I found that we had missed the central use of *of*. We'd got all kinds of marginal uses, we had *die of the plague*, and we had the im-, even, we even had the American construction: *It's a quarter of three*, which means 'it's a quarter to three', and all kinds of thing[s]. And we had the posse-, we had *the love of God*, *the shooting of the hunters*, *the shooting of the elephants*, all those things from, hot off the linguistic press. And we had some had some possessives of various kind[s]. But we didn't have *the colour of her hair*, that very elementary use of *of* to associate one thing to another in some general way: *the colour of her hair*, *the size of the room* and so on, which turned out to be by far the most common use of *of* in the files of the Survey of English Usage. Fortunately, the *Longman Dictionary of Contemporary English* had not yet gone to press, and we were able to put this crucial sense of *of* into the dictionary in time so it is there.

Well, as I had done the first draft of *of*, I was about to take the honourable way out for a lexicographer, which is suicide in such cases [laughter]. How could I have been so stupid as to leave out the most common use of *of*, especially, when, like all good lexicographers, I had done my share of research, or, as Tom Lehrer calls it, plagiarism? I'd looked at all the other dictionaries, hadn't I? But I'd missed this central sense of *of*. Well, I went back, before, you know, killing myself, I went back, and I looked at all the other dictionaries I'd examined, and they didn't have it either. I'm sure the *Oxford English Dictionary* had it, but I didn't look at that. But all the other

learners' dictionaries, the other small ((or)) middle-size dictionaries had simply not got this common sense of *of*.

And why? Well, presumably, precisely because it is so common. It is so colourless, so common, so central that it'd been missed. And the lexicographers had focussed on the unusual, the singular, the spare, the strange, as Hopkins says in one of his poems. And had missed the common, the central, the uncontroversial. But thirty seconds looking at the Survey of English Usage, which, as I said before, must take every single example of *of* in all our texts, thirty seconds, and the central use as well as the marginal uses became evident.

And, of course, we can also use the Survey corpus as actual dictionary examples. We can u-, the Survey material can be used also as e-, as exemplification in coursebooks and grammars, and there is a lot of exemplification from the Survey in grammars like the new *Comprehensive Grammar of the English Language* by Quirk et al.

The Survey, however, does not provide a-, the answers to all your questions about English, nor can a corpus in principle. Pa-, that's partly because of our theory-neutral system of analysis. Thus, students sometimes come to me and say: "We'd like to see your file of questions, please." Perfectly reasonable request. Or: "We'd like to see your file of orders, or mands as they're someti- [=sometimes] called, or suasions. T-, Z show me your file of suasion, please." Well, we have no file of suasion. We haven't even got a file of questions, as you will see if you look at our grammatical categories.

So. Sometimes, however, you can get round the problem by ingenious methods of search. You can't find all our qu-, questions in the Survey, but you can look up all our question marks, because, as marks of punctuation, they are all saved. Therefore, you can probably find many, though not a-, all, of the questions in our written material, even if you can't find them so easily in the spoken material, one way of ((looking into it)).

Now, like any other corpus, the Su-, or corpus concordance, the Survey suffers from two major kinds of problem. One is information

glut, and one is information famine. And we have ways of dealing with both. There is a famous story about someone who came to a psychiatrist, and said that his problem was that he was interested in pancakes [laughter]. And the psychiatrist said: "What's wrong with that? I like pancakes myself". And the patient said: "Oh, you must come and visit my flat then. I have several rooms full of them" [laughter]. That is one of the problems of the Survey of English Usage. You may, for example, be very interested in the English definite article *the*. Well, you must come to the Survey of English Usage, [laughter] 'cos we've got rooms full. We haven't got rooms full, but we've got about five filing cabinets[8] full of slips, because we record every instance of *the*. If you're interested in *t-o*, we've got lots of *t-o*s, as well. Drawers and drawers full of them. Which is wonderful, except that it's very difficult to process so much information for any project of limited scope.

So how do we get round that? We get round information glut by the process that we call 'Second Stage analysis'. A more delicate stage of analysis. So, for example, in the case of *t-o* in English, there is a very obvious thing that we should do: we should take all our hundreds of, or our thousands of *t-o*s, and divide them into those which are the marker of the infinitive: *I want to go*, and all the other *to*s, which are ordinary prepositions: *go to the circus* as opposed to *want to go*. In that way, we will immediately reduce the problem for scholars who want to look at *to*. They can then decide whether they want to look at *to*, the marker of the infinitive, or *to*, the preposition. Even so, there'll still be a lot of material, but it will be in more manageable fashion. And we can make further refinements of analysis, if and when we have time and resources.

At the opposite extreme we have the problem of information famine. At the mo-, we have, at the moment, as I say, about one or two examples of *albeit*, the function word, in our files. Not ten draw-

---

[8] exact number uncertain (lecturer's written note)

ersful, I can assure you. So that's clearly not enough for somebody to write a doctoral dissertation on *albeit*.

And there are many other questions that, that even a relatively generous Survey file cannot answer by itself. For example, the, the adverb *utterly* in English has some very strange features. It tends to go with negative things. You can say: *That argument is utterly wrong*. But you are unlikely to say: *That argument is utterly right*. Although you can say: *The argument is completely wrong. The argument is completely right*. On the other hand, it's not as simple as that, because you can say things like: *The argument, your, the argument is utterly, the, the, the, the programme last night was utterly marvellous, utterly fantastic, utterly divine*, which are all highly positive, even though they are a bit effusive. So how can you test the semantics of *utterly*, and what it combines with? You will probably not get enough material from the corpus to test it out, and so the corpus material must be supplemented by elicitation tests, simple psycholinguistic experiments, of which I have a, a handout[9] here. This is the full form, of which, as usual, I've got about ten, and then, there are supplementary sheets, which, unfortunately, give only some, but not all, of the types of tests we have.

Table 4

CORPUS MATERIAL

(a) *print*

> (...) This conclusion was confirmed and extended by Wieland and his co-workers (1935), who obtained 2- and 4-methyldiphenyl from the decomposition of (...)

[9] Table 4 (see pages 32-34)

(b) *manuscript*

(...) Not so rich & fatty. So I – gred – griddled bacon & tomatoes for breakfast & they were jolly good to – Coo! (...)

(c) *speech*

(...) a m /that's where – it's cŏming from#a#m# and /Marsh tŏld [/him#]# ə: that so m/far as (...)

ELICITATION TESTS

(d) *degree of compliance – putative problem in test sentence*

They are /putting the light in the far cŏrner off # — To Past

(e) *degree of compliance – putative problem in target sentence*

They will /probably stay làte # — To Question

(f) *selection*

He /dared to answer me bàck # — To Negative

(g) *forced-choice selection*

As a boy, I often —— of eagles.

Last night, I —— of eagles.

} Fill one blank with *dreamt* and the other with *dreamed*

(h) *word placement*

HE CAN NOT DRIVE A CAR — Write down the sentence using *probably* with it

(i) *composition*

They /badly	complete the sentence

(j) *evaluation*

I was /sat opposite by a stránger #	Judge; Yes / ? / No *or* on 5-point scale from 'completely unacceptable' to 'perfectly OK'

(k) *similarity in meaning*

The /book was unfortunately dìfficult # Un/fŏrtunately#the /book was dìfficult #	Judge: Very similar / ? / Very different

(l) *preference – rating*

They needn't see a lawyer. They don't need to see a lawyer.	Judge as for (j), but in pairs

(m) preference – ranking

He doesn't have a car. ☐ He hasn't a car. ☐ He hasn't got a car. ☐	Mark with order of preference

(n) *frequency*

We recommend that he pay full tuition. We recommend that he should pay full tuition.	Judge: on 5-point scale from 'very rare' to 'very frequent'

I heartily recommend these elicitation tests to anyone interested in the investigation of language, spoken or written, and of items that are disputed or where there is variable usage, or where the boundaries or the constraints are slightly vague and difficult to determine. We keep inventing new kinds of elicitation tests, but this [list] gives, oops, this is an odd one [about a sheet of the handout], seems to be upside down, but anyway.

For example, I, some of you may loo-, have, well, you can look at, some of you may have test type (i): composition, although I'm afraid that not all of you do. This is a very simple kind of test, the completion test: *They/badly*, to be completed by some kind of phrase, which could be *They badly need a, a drink*, *They badly need a-, another xerox machine*, etcetera. But it's entirely possible that *badly*, that there are some constraints on the kind of verb phrase that can be intensified by *badly*. And that we can find simply by getting people to complete the sentence. That's very elementary.

Similarly, if you look at something that you probably have got, (g): the forced-choice test. The, the forced-choice test is one of my favourite tests. It's very easy to do, it produces very good results if you observe certain methodological constraints. It's, it's marvellous for items of disputed usage. Here the illustration is to compare *dreamt* and *dreamed* as past forms of the English verb *dream*. Very simple. You have two sentences, with each of, which have a blank: *As a boy I often -blah- of eagles. Last night I -blah- of eagles*, and you have to fill one blank with *dreamt* and one with *dreamed*; even if you can use both in both spaces, even if you can use neither in both spaces, even if, even if one of them seems impossible, you still have to use both. If you do, some very interesting things will emerge. What will emerge is, is that *dreamt* will be favoured in the second sentence: *Last night I dreamt of eagles*, and *dreamed* will be favoured, I think, in the first sentence: *As a boy I often dreamed of eagles*, suggesting that *dreamed*, which is, after all, a longer word because it has an, ends in a voiced consonant, which is longer, and

has a longer vowel in typical pronunciation, seems to go rather well with sentences that express duration of the action, or repetition of the action, and *dreamt*, which is short and to the point, goes well with things that happen only once.

Well, that's interesting. But there's something else that emerges from this particular test. In principle, *dreamed* is found in both British and American English, *dreamt* is found far more in British English than in American English. In fact, some people might even say that *dreamt* is British English and not American English. But this sort of test has been carried out on American speakers as well as on British speakers. Now, American speakers may not use *dreamt* spontaneously, but in a forced choice test they have to use it [laughter], and, you know, we, we tie them to a chair, and administer electric shocks if they don't comply. And if they have to use it, it turns out that they use, that they make a distinction between *dreamt* and *dreamed* which is pretty much the same as the distinction made by British speakers. Which suggests that many differences between varieties of the same language are surface differences having to do with relative frequency, based on an underlying lexico-grammatical s-, system that turns out to be surprisingly similar in the two varieties, or, that're being compared.

Well, I will just say two other things about the Survey of the many that I could. One is that we are trying to make Survey material available to as many people as need it. The only way to get the whole lot is to come to University College, which, I hope, you will all do immediately: there are planes leaving regularly. However, the individual Survey texts, each 5,000 words long, as you recall, can be bought. We can cop-, we can xerox them and sell them to you for a modest sum. Ho-, furthermore, a book version of our spontaneous conversations, S.1 to S.3, is now available, edited by Jan Svartvik and Randolph Quirk, published by, published by some Swedes [laughs], by, [audience trying to help] (that) name do-, y-, doesn't sound, yes, anyway, tha-, it is available in book form, as *A Corpus of Spoken*

English,[10] in a s-, with a simplified transcription and no grammatical analysis, but if you've never seen what real spoken English looks like when in print, this is, this is the way to, to look at it. And I've long wanted to do a comparison between real conversation and the dialogue in novels and plays, by the way: it'd be absolutely fascinating. We are, now there is a computer tape of many, but not all of our s-, S-texts, which is available from the Norwegian Computing Centre for the Humanities, as the London-Lund corpus, and we are now, as I say, pru-, putting the whole of our corpus on computers that IBM have put at our disposal. We are producing a new, new methods of analysing the material, using semi-automatic tagging procedures for individual items, lexical items, and for syntactic structures. That is now going on, and when it is further advanced the computer tape may, or may not, be made available to the general public. That is now under discussion.

To summarise: the Survey is a valu-, is valuable both as a corpus, that is, simply as, as a, as a collection of texts, and as a concordance, that is, as an analysed collection of texts taking all tokens of e-, all tokens of each category.

The Survey was not founded to answer specific questions. What I mean is, for example, there is a lot of divided usages, you'll see from one of the handouts,[11] about how you negate the verb *have*. Do you say *don't have* or *haven't got* or something of that sort? There is a lot of trouble about the negative of *used to*. Is it *usedn't to*, *didn't use to* or what? The Su-, nobody, I think, has ever given the Survey of English Usage a large sum of money to investigate a specific problem. Individual researchers have had grants to work on specific problems.

The Survey was funded and continues to be funded to provide a lot of analysed data which can be used by anybody to answer a wide

---

[10] *A Corpus of English Conversation*. Lund: Gleerup 1980.

[11] Test type (m) in Table 4 (page 34)

variety of questions. That is why it has survived so long, and that is why it goes on surviving. You will find in it information relevant to all the problems I've been discussing, and to many others. And one of the most important things is that the Survey isn't even in existence [just] to answer questions. It is in existence as much to raise questions as to answer them. If you spend a day at the Survey, you may find the answers to some questions, but I hope that you will also find some surprises. You will find some problems about English, maybe even about language in general, that you didn't think existed. You'll find problems where you thought there were only simple, straightforward, general statements. So the Survey exists not only to bring order into chaos, of the chaos of dis-, but also to bring chaos into the tidy order of our theoretical preconceptions.

And over the years, as I say, the Survey's own theoretical conceptions have been influenced by changes in the language, by ch-, for example, the rise of new vocative forms, which I won't go into, by changes in the mode of production and the mode of collection and the mode of diffusion of texts, by changes in our minimal theoretical framework forced upon us ourselves by the data that we ourselves have collected. And in all these ways the Survey has reflected the age or ages in which it has been going on [laughter as he slows down]. Well, thank you.

Appendix

Selected books using Survey of English Usage material:

Aarts, F. & Jan (1982): *English Syntactic Structure*. Oxford: Pergamon.

Aarts, Jan & Willem Meijs (1984-1986): *Corpus Linguistics I-II*. Amsterdam: Rodopi.

Bald, W. D. & Robert Ilson (eds.) (1977): *Studies in English Usage: The Resources of a Present-Day English Corpus for Linguistic Analysis*. Frankfurt: Lang.

Biber, D. (1988): *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Crystal, David & Derek Davy (1969): *Investigating English Style*. London: Longman.

Crystal, David & Derek Davy (1975): *Advanced Conversational English*. London: Longman.

Crystal, David & Randolph Quirk (1964): *Systems of prosodic and Paralinguistic Features in English*. The Hague: Mouton.

Greenbaum, Sidney (1969): *Studies in English Adverbial Usage*. London: Longman.

Greenbaum, Sidney, Geoffrey Leech & Jan Svartvik (eds.) (1980): *Studies in English Linguistics*. London: Longman.

Greenbaum, Sidney & Randolph Quirk (1970): *Elicitation Experiments in English: Linguistic Studies in Use and Attitude*. London: Longman.

Leech, Geoffrey & Jan Svartvik (1975): *A Communicative Grammar of English*. London: Longman.

Longman Dictionary of Contemporary English (1978[1]). Paul Procter (editor-in-chief), Robert F. Ilson (managing editor). London: Longman.

Oreström, B. (1983): *Turn-taking in English Conversation*. Lund: Gleerup.

Quirk, Randolph (1968): *The Use of English*. London: Longman. 2nd edition.

Quirk, Randolph (1974): *The Linguist and the English Language*. London: Arnold.

Quirk, Randolph & G. Stein (1990): *English in Use*. London: Longman.

Quirk, Randolph & Sidney Greenbaum (1973): *A University Grammar of English*. London: Longman.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik (1972): *A Grammar of Contemporary English*. London: Longman.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik (1985): *A Comprehensive Grammar of the English Language*. London: Longman.

Quirk, Randolph & Jan Svartvik (1966): *Investigating Linguistic Acceptability*. The Hague: Mouton.

Svartvik, Jan (1966): *On Voice in the English Verb*. The Hague: Mouton.

Svartvik, Jan (ed.) (1990): *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund University Press.

Svartvik, Jan & Randolph Quirk (eds.) (1980): *A Corpus of English Conversation*. Lund: Gleerup/Liber.

Ára: 50,– Ft