ED 322 776 FL 018 792

AUTHOR Zora, Subhi; Johns-Lewis, Catherine

TITLE Lexical Density in Interview and Conversation.

PUB DATE 89

NOTE 14p.; In: York Papers in Linguistics 14; see FL 018 786.

PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS College Students; Comparative Analysis; Foreign Countries; Higher Education; *Interpersonal Communication; *Interviews; Language Research; *Language Styles; Linguistic Theory; *Vocabulary

ABSTRACT

A study examined lexical density in interviews and conversation with the same subjects. Sixteen undergraduate and graduate students, members of religious, political, and/or cultural societies at Aston University (England), were interviewed in pairs by the university chaplain, who knew them. Immediately following each interview, the chaplain left the two students to converse freely. The pairs were already friends. Analysis of the students' language use found that while the lexical density was different in the two situations for graduate students, with the density higher for the interview, under raduate students performed comparably on this variable in both settings. Accommodation theory appears to be applicable to the lexical level of linguistic control. Further investigation of the differences in skill and/or sensitivity at the lexical level is recommended. Eight possible sources of variation are identified, including: bases for calculating lexical density; expected interruption and length of speaking turn; function of component units of text; self-consciousness/self-monitoring; personal attributes; group attributes; planning time; and topic. It is suggested that these variables be controlled in future research. (MSE)

LEXICAL DENSITY IN INTERVIEW AND CONVERSATION

Subhi Zora and Catherine Johns-Lewis

LEXICAL DENSITY IN INTERVIEW AND CONVERSATION[*]

Subhi Zora and Catherine Johns-Lewis

Aston University

1. Introduction

Lexis is a potential indicator text type, in that variation in lexical frequency, lexical complexity and lexical relations can differentiate between types of spoken or written discourse (Ure, 1971; Halliday, 1985; Stubbs, 1986). This paper investigates lexical density (LD) in two varieties of spoken discourse: interview (INT) and conversation (CON), the data being produced by the same subjects. The hypothesis explored is that at least one source of LD variation is personal maturity. However, since inter-individual variation cannot be explained entirely on the basis of this factor, other socio-psychological parameters are clearly relevant.

2. Lexical vs. grammatical items: definitional comments

For Lyons (1986), Robins (1964) and Palmer (1976) lexical items are the major content words, which fall into four grammatical categories: Nouns, Adjectives, Adverbs and Main Verbs. Grammatical items (or function words) serve to express relations between content words, and include: Auxiliary Verbs, Modals, Pronouns, Prepositions, Determiners and Conjunctions.

Sorting words into lexical (L) as opposed to grammatical (G) sets is of course not entirely straightforward. In the so-called phrasal verbs,

[*] Authors' correspondence address: Department of Modern Languages, Aston University, Aston Triangle, Birmingham B4 7ET.

York Papers in Linguistics 14 (1989) 89-100

the status of the preposition or particle element is sometimes difficult to determine. For example in:

(1) She made up her face (from Halliday 1966: 153)
(2) She made up her story
(3) They made up and kissed
(4) She made up the hill at speed

the grammatical object in (1) is optional (as in *She made up swiftly*), but not so in (2). The *up* in (1) would therefore appear to be more of an adverbial particle than a preposition, and is therefore directly comparable with the particle status of *up* in (3). In (4), *up* is of course prepositional. The point is that:

(5) They made up

is ambiguous as between meaning (1) and meaning (3). The implication is that in a sensitive analysis, the grammatical status of *up*, can only be resolved by reference to the lexical context. For the purpose of statistical analysis of lexical versus grammatical words, the distinction between the subcategorisation of *make up* has to be ignored, in favour of a cruder classification. In our calculations and for the purposes of this paper, we have followed both Ure (1971) and Stubbs (1986) in considering such phrasal verbs as *make up* as consisting of two words, one lexical word *make* and one grammatical *up*.

3. Preliminaries to Research on LD

Previous research of LD has shown that the concept of 'density' (i.e. ratio of L to G items within a text) can allow texts to be ranked in relation to each other. In very general terms, the ratio of L to G items will show how lexically dense one text is as compared with another.

Lexical density is a property of text, to be calculated in terms of the frequency of L and G items. Information density is a property of processing, for which there is no valid absolute statistic, in that the same text, with its definable LD, will represent different information processing loads for different readers. Space does not permit this point to be expanded. The reader is referred to information processing theories (e.g. Schank, 1975).

Two approaches have been used by researchers to arrive at the G:L ratio in the literature on spoken and written discourse. The first approach is manual, whereby the status of all words in a text is noted by the analyst, after which percentages are worked out (see for example Ure 1971). The second approach is automatic and depends mainly on computer programmes like the one devised by Stubbs (1986), which was designed to run on the London-Lund corpus of spoken English[1]. The manual approach has a greater degree of accuracy since each problem is dealt with by the human linguist in its real context. However, the amount of text processable is presumably limited. Automatic analysis based on tailor-made software, though efficient and reliable to a great extent, is not without problems. In addition to the problem of phrasal verbs mentioned earlier, other types of problem can also arise, some of which have been reported by Stubbs (1986), which no algorithm can resolve. One such problem is that some of the modal verbs such as *can* and *will* can also occur as main verbs, or as nouns in certain contexts. Auxiliary verbs such as *be*, *have* and *do* can also be G or L according to the grammatical contexts in which they are used. Stubbs (1986) solved such problems in his program by building into it a routine to deal with potentially ambiguous words which are categorised according to their context in running text.

There is of course a more general problem in word classification. What one researcher counts as lexical, another will classify as grammatical. Stubbs (1986), for example, lists *be* as lexical or grammatical. Ure, on the other hand, commenting on her (1971) results, counts it as

[1] See Svartvik, J. et al. (1982) for a detailed description of this project of spoken English.

grammatical, even when it has a more lexical function (as in *if you don't be good...*), (personal communication).[2]

Stubbs (1986) provides a useful list of G words, non-G words being, by implication L words. However, the list may not be exhaustive. While, for example, *anything* and *sometimes* are includec, *anyone* and *something* are not.

4. Research on LD in types of discourse:

Ure (1971) manually calculated LD in 34 spoken texts and 30 written texts comprising approximately 21,000 words each. The former texts, all except two, have a strong tendency to have an LD of less **than** (40%), whereas the written texts, all except two, have a strong tendency to have an LD of **greater than** (40%). Although these results are suggestive, they are not conclusive, since different subjects produced the spoken and the written data. This is an important source of variation as shown by Beaman (1984) and Farag (1986). The literature on spoken versus written language is considerable and will not be gone into here.

Stubbs (1986), adopting the program mentioned above to analyse six spoken sub-texts of the London-Lund corpus (op. cit.), which represents recordings of highly educated informants, mostly academics, found a significantly higher LD than is reported by Ure. Stubbs' computer calculations show an LD ranging between (44%) and (56%). He relates the difference between his results and Ure's to the different study methods used in the calculations and the nature of the corpora studied. He also mentions the level of respect, which, as we shall see later, is confirmed by our results as well.

Hasan (1988, forthcoming) compares LD in native and non-native speaker speech in five types of formal and informal types of spoken dis-

[2] We are very much indebted to J. Ure for her invaluable comments on her 1971 work and for comments on an early version of this paper.

course. He reports his formal native-speaker interviews to have an LD of (47.02%) and informal conversation of (42.48%). These results are similar to the results of the present investigation as we shall see below.

5. The Subjects

16 subjects (6 postgraduates and 10 undergraduates), who were all members of religious, political and/or cultural societies at the Guild of Students, Aston University, were interviewed by a university chaplain, who knew them all on a personal basis or through religious contact. The interview took the form of a review of personal development over the previous 12 months, the chaplain in each case acting as 'elicitor of insights'. Subjects were recorded being interviewed in pairs using a UHER 4000 REPORT recorder with the microphone about 1 metre from participants in a quiet environment. Immediately following the interview, the chaplain withdrew, leaving the two subjects to chat freely. The pairing of the subjects was elective: each pair representing a 'close friend'.

6. Analysis

Calculations of the L:G ratio were first done manually then computationally. In the second method, two simple computer programs[3] were used to identify all L and G items. The final calculations represent an adjustment of the computer programs so as to take account of ambiguous classification. In all ambiguous cases, context was the basis of the decision. The results and statistical tests of significance[4] are presented below.

[3] The two computer programs employed are 'ALFSORT' and 'FREQSORT' both devised by Professor Frank Knowles at the Department of Modern Languages, Aston University.

[4] 'Wilcoxon's Matched Ranked Pairs Test' was used to test the statistical significance of the results. Details of this test are to be found in Meddis (1975).

Before presenting LD results for INT and CON, two factors must be mentioned which could affect results: repetition and interview input. In order to differentiate between the output of subjects and the output of the interviewer, Table 1 figures include interviewer output and repetitions, while Table 2 figures exclude interviewer output and repetitions. Each pair of subjects is indentfied as A, B, C, etc. A, B and C pairs are postgraduates, pairs D to H being undergraduates.

TABLE 1
Overall Lexical Density in INT and CON

Pairs	LD in INT	LD in CON
A	48.2%	46.9%
B	47.2%	44.4%
C	50.4%	47.4%
D	46.9%	44.3%
E	47.4%	47.6%
F	45.3%	47.3%
G	48.7%	47.6%
H	43.7%	46.3%
Mean	47.2%	46.5%
SD	2.052	1.379

Table 1 above shows a slightly higher mean percentage of L words in the interview situation. However, the difference is not statistically significant (Wilcoxon Signed Rank Text statistic 10.500, p< 0.147). The higher Standard Deviation in the interview data reflects the greater spread of LD values than in conversation.

In order to assess the influence of repetitions and interviewer input, the L percentages were recalculated excluding these. Table 2 presents corrected figures. It shows that, excluding interviewer speech and repetition from the calculations, again LD is somewhat higher in INT than CON but the difference is not statistically significant (Wilcoxon Signed Rank Test statistic 8.000, significance level 0.081). Again, we find that the higher Standard D 'iation of the interview data reflects the fact 'hat there is a greater spread of LD values. In other words, it is a less internally consistent set of figures than the conversation set.

TABLE 2
Lexical Density in INT and CON
Excluding Repetitions and Interviewer Speech

Pairs	L in INT	L in CON
A	49.8%	47.9%
B	48.5%	45.7%
C	51.1%	47.6%
D	48.7%	44.9%
E	48.2%	47.9%
F	45.7%	47.4%
G	48.2%	47.7%
H	44%	46.6%
Mean	48.025%	46.962%
SD	2.235	1.126

A comparison of mean percentage values for postgraduates as opposed to undergraduates is revealing.

	INT	CON
Postgrad	49.8%	47.0%
Undergrad	46.9%	46.9%

Although the numbers are too small for valid statistical testing, there would appear to be some evidence that:

(1) undergraduates do not differ from postgraduates, in terms of lexical density, in the conversational setting

(2) undergraduates, who are less mature and have received a shorter period of higher education, do not increase LD in the formal

interview, whereas postgraduates do. This possibility requires further work.

One aspect of the results above so far not discussed is that, as Ure (1971) found, the absolute LD value varies from one participant to another, in the same speaking task.

TABLE 3
Lexical Density in Subjects' speech in INT and CON

Participants	LD in INT	LD in CON
HC	48.1%	47.0%
KW	50.4%	47.8%
JH	48.7%	46.8%
BG	48.4%	44.2%
RH	52.0%	46.4%
DD	50.4%	49.1%
RF	50.0%	45.4%
PM	48.0%	41.3%
HH	47.4%	47.3%
KSH	48.7%	48.5%
AM	46.9%	48.3%
CB	44.8%	45.9%
JC	50.6%	47.9%
GM	46.5%	47.5%
AB	44.1%	46.9%
KS	43.8%	46.3%
Mean	48.0%	46.9%
SD	2.235	1.126

The differences between individual speakers are in general as great as the differences between the two speaking tasks. It is also worth pointing out that the direction of difference is not consistent. There are 4 individuals for whom CON has a higher LD than INT (CB, GM, AB, KS); there are 3 individuals who produce the same or virtually the same LD in INT and CON (DD, HK, KSH); and there are 9 individuals for whom there is a clear step up in LD in INT compared with CON (HG, KW, JH, BG, RH, RF, PM, AM, JC). Thus, almost half the speakers manifest a trend which is not in agreement with the trend established by averaging across the whole population. If lexical density is affected by maturity and educational level, further work paying attention to the output of individuals will be required.

7. Conclusion

The general conclusion is that in the present study, which is an attempt to have the same speakers perform different speaking tasks in a controlled situation, lexical density does not differentiate between discourse modes in a global way. Rather, it differentiates between interview and conversation for the postgraduates analysed. Undergraduates, on the other hand, perform comparably, in terms of lexical density, in both the interview and the conversational setting. Since the population examined is a) small and b) not evenly balanced (as between undergraduates and postgraduates), it is premature to conclude that an absolute statistic for the lexical density of undergraduates and postgraduates can be produced. What is interesting, and worth pursuing further, is the differential between the two groups in skill and/or sensitivity at the lexical level. It would appear that postgraduates, who are more mature and have longer exposure to higher education, adjust their lexical density to match some perceived characteristic of the interview situation. Postgraduates may be more able to compete on an equal footing with the interviewer, and this ability may derive in part from a perception that their own status is close to that of the interviewer. Essentially, what is being suggested is an application of 'accommodation theory', which is well known in social psychology, to the lexical level of linguistic control, as an explanation for the rise in lexical density in the interview situation. The interviewer's drop in lexical density can be seen as a conciliatory gesture, metaphorically the opposite of a claim to status; and this in turn facilitates the closure of the status gap of which the postgraduates are able to make use.

8. General discussion

The LD levels of the spoken data analysed in this work are considerably higher than those reported by Ure (1971) for her spoken data and are generally closer to those reported by Stubbs (1986) and Hasan (1988). Even the lowest percentage obtained is higher than the highest in Ure's spoken data where percentages range from 23.9% (assembling Angel Chimes) to 43.2% (radio sports commentary). In Stubbs (1986), the range is from 44% (business telephone conversations) to 56% (radio state funeral commentary).

The question we would like to now ask is: why do different researchers report very different percentages for apparently identical speaking tasks? (Compare Stubbs's 54% for radio cricket commentary with Ure's 43.2% for radio football commentary; or Ure's 'Life' discussion among students (35.2%) with the figure for conversation between students (46.9%) in the present study).

There may be at least eight sources of variation:-

(1) **basis for calculating LD:** i.e. differences in allocating items to lexical as opposed to grammatical classes.
(2) **expected interruption and length of speaking turn:** longer monologic texts predisposing speakers to higher LD (see figures in Stubbs (1986) and Ure (1971) where spoken texts with higher LD are monologues, such as sermons, House of Commons debates, radio commentaries, or lectures).
(3) **function of component units of text.** In the present study, when units with narrative, informative, inquisitive, argumentative or responsive functions are compared, the hierarchy of LD is informative>narrative>inquisitive>negation/hesitation/hedging. The LD (43.9%) of interviewer speech, which is inquisitive, repetitious, full of hedges, and hesitant is lower than the mean LD (47.8%) of interviewee speech.
(4) **self-consciousness/self-monitoring.** Compare Ure's figures for lecture (39.6%) and recorded language laboratory instructions

(40.9%) with the mean 46.0% in interview, 46.9% in conversation, obtained in the present study.

(5) **personal attribute:** maturity, educational level, confidence. Stubbs (1986) comments that the high LD obtained in his study of the London-Lund corpus could have been the product of the high educational level of the speakers. Similarly, Ure (1971) talks of the influence of the previous experience, skill and education on the performance of her subjects.

(6) **group attributes:** age, sex, educational level, etc. In the present study, undergraduates produce lower LD in the interview situation than postgraduates. It should be noted that group attributes may not always be distinguishable from personal attributes.

(7) **planning time.** Both Ure (1971) and Stubbs (1986) mention this as distinguishing between spoken and written production, and it may also contribute to the monitored/unmonitored distinction.

(8) **topic.** Stubbs presents a different LD for state funeral commentary (56%) as opposed to radio cricket commentary (54%). The same 'genre' with different topic and presumably different textual sub-functions can manifest different LD levels.

It is clearly desirable that all eight factors should be controlled in experimental studies of lexical density, although the difficulties of doing so are not underestimated. Ure (1971) for example has two almost directly comparable texts: a spoken text (LD 32/2%) 'How to repot a plant' and a written text (LD 47.1) 'Planting and soil'. It may be difficult to obtain a direct spoken counterpart of a written text; or, indeed, there may be no direct spoken counterpart. (What would be the spoken counterpart of a television news text, which is normally read aloud from a teletext machine?)

There is scope for applying algorithms such as the one developed by Stubbs (1986) to data as wide-ranging as Ure's (1971), but designed in such a way as to ensure that the same subjects produce contrasted text types, on the same topic. Until we know more about the sources of variation in lexical density, explanation of the functions of variation in lexical density will remain tentative.

REFERENCES

Beaman, K. 1984. Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In D Tannen, ed. *Coherence in Spoken and Written Discourse*. Norwood, New Jersey: Ablex. 45-80.

Farag, S. 1986. *A Linguistic Analysis of Spoken and Written Narrative Discourse*. Unpublished PhD Thesis. Aston University.

Halliday, M.A.K. 1986. Lexis as a linguistic level. In Bazell C. B. et al. eds. *In Memory of J R Firth*. London: Longman. 148-162.

Halliday, M.A.K. *Spoken and Written Language*. Victoria: Deakin University.

Hasan, A.S. 1988. Forthcoming. *Variation in Spoken Discourse in and Beyond the English Foreign Language Classroom*. PhD Thesis. Aston University.

Lyons, J. 1968. *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.

Meddis, R. 1975. *Statistical Book for Non-Statisticians* London: McGraw Hill.

Palmer, F.R. 1976. *Semantics*. Cambridge: Cambridge University Press.

Robins, R.H. 1964. *General Linguistics: An Introductory Survey*. London: Longman.

Schank, R.K. 1975. *Conceptual Framework Processing*. New York: Elsevier Press.

Sinclair, J. 1966. Beginning the study of lexis. In C.B.Bazell. et al. eds. *In Memory of J R Firth*. London: Longman. 410-430.

Stubbs, M. 1986. Lexical density: A computational technique and some findings. In M.Coulthard. ed. *Talking about Text*. Birmingham, University of Birmingham: English Language Research. 27-48.

Svartvik, J. M, Eeg-Olofsson, O., Forsheden, B., Orestrom and Thavenius, C. 1982. *Survey of Spoken English*. Research Report 1975-1981. Lund: Gleerup.

Ure, J. 1971. Lexical density and register differentiation. In Perren, G.E. and Trim, J.L.E. eds. *Applications of Linguistics*. Cambridge: Cambridge University Press. 443-452.