Michael Abecassis
Oxford University

Saliency and frequency in a corpus of 1930's French films

For this research, a corpus of French films (recorded on videocassette) dating from the 1930s has been assembled: this provides interesting and previously unexploited evidence concerning Parisian vernacular speech at that time. My film corpus comprises five black and white films: *Hôtel du Nord* (1938), *Fric-frac* (1939), *Circonstances atténuantes* (1939), *Le Jour se lève* (1939), *La Règle du jeu* (1939). In the five films I investigated, the script-writers clearly blur all social distinctions. I end up with a caricatural picture of Paris society divided into two social groups: the lower group on the one hand and the upper group on the other. I chose these films first because of their lasting popularity: they are some of the most famous films of the 1930s. Above all, however, I thought they were representative of the most stereotypical Parisian sociolect of that period. In this article, I intend first of all to explore the differences in lexical behaviour between the two social groups of characters. For this purpose, I will apply statistical methods developed in corpus linguistics (Butler 1985, Scott 1996) to see whether the lexical behaviour of the two sub-groups can be seen to differ according to "saliency" and "frequency". In the second part of my study, I intend to evaluate the extent to which the dialogues in my films which are artificial and do not constitute natural language approximate naturally-occurring conversations.

1.0. Saliency

By "saliency", I mean all the words that stand out statistically when one subcorpus is compared to another subcorpus or to the totality of the corpus. For this exercise, I used Mike

Scott's word -list (1994) programme, which can produce a word-list analysis comparing two texts or two corpora. In comparing two texts, say the lower-group speech and the upper-group speech in *Fric-frac*, the programme enables us to identify the salient words of one subcorpus relative to their occurrences in the other. A given word whose frequency in a source text is statistically greater or smaller than its frequency in a larger word list based on a reference corpus is called a key word (Scott 1996). I would expect, for example, the article "le" to be of a high frequency in any French text but it is not necessarily salient. It is only salient if in most texts one finds, say, 7% of "le", but in a particular text one gets 15%. It would also be salient, but this time negatively, if one only found 1% of "le" in a particular text compared to 7% in the whole corpus. So, saliency has to do with being noticeably different, statistically speaking. The computer calculates the frequency of each lexeme in each subcorpus and estimates the statistical significance of any differences. Saliency is assessed by taking into account the frequency of one lexeme in a text in comparison with its occurrences in another text or in the main corpus.

To begin with, I put together all the cues of the main upper-group speakers and all those of the main lower-group speakers to constitute in each film two contrasting texts: Fric 1 and Fric 2, Le Jour 1 and Le Jour 2 etc. I then assembled all the texts of each film to constitute two main subcorpora named Subcorpus 1 and Subcorpus 2. Subcorpus 1 is the totality of the speeches of the main upper-group speakers (Geneviève, Madame, Marcel, Monsieur, Pierre, Renée, Renée, Robert, Valentin) in the five films put together. Subcorpus 2 is the totality of the speeches of the main lower-group speakers (Bouic, Clara, Edmond, François, Françoise, Jo, Loulou, Marceau, Marie, Raymonde) in the five films put together. In addition, I created a film corpus by putting the five films together. Subcorpus 3 is the totality of all the speakers' speeches in all films.

Subcorpus 1 (upper group)	Subcorpus 2 (lower group)
Fric 1	*Fric* 2
Circonstances 1	*Circonstances* 2
Jour 1	*Jour* 2
Règle 1	*Règle* 2
Hôtel 1	*Hôtel* 2
Subcorpus 3 (all characters all films combined)	

Table 1: different subcorpora

I first established for each film two word lists to compare the lower-group speeches and upper-group speeches in each film with the film corpus as a whole:

Upper group	Lower group
Word list 1a: *Fric* 1 versus Subcorpus 3	Word list 1b: *Fric* 2 versus Subcorpus 3
Word list 2a: *Circonstances* 1 versus Subcorpus 3	Word list 2b: *Circonstances* 2 versus Subcorpus 3
Word list 3a: *Jour* 1 versus Subcorpus 3	Word list 3b: *Jour* 2 versus Subcorpus 3
Word list 4a: *Règle* 1 versus Subcorpus 3	Word list 4b: *Règle* 2 versus Subcorpus 3
Word list 5a: *Hôtel* 1 versus Subcorpus 3	Word list 5b: *Hôtel* 2 versus Subcorpus 3

Word list 6 compares in a final stage Subcorpus 1 to Subcorpus 2.

Each word list gives us a ranked list going from the most salient lexemes to the least. For each lexeme, the programme measured a chi-square score to evaluate whether the frequency of a particular lexeme is statistically significant across the two subcorpora. A low chi-square score indicates that the frequency of a lexeme is not high enough to be significant. A high chi-square score, on the other hand, suggests that the proportion of *tokens* of a lexeme in one subcorpus, in comparison with another subcorpus or the corpus as a whole, is great enough not to be random. I will concentrate on the lexemes that obtain a high chi-square score. It is hoped that this programme will help us to distinguish a pattern of linkage and difference between Subcorpus 1 and Subcorpus 2.

1.1. Saliency between the upper/lower-group subcorpus within each film and Subcorpus 3 (word lists 1a-5b)

Scott's word list programme enabled us to investigate which words were salient in the upper-group and lower-group subcorpora of each film in comparison with the whole film corpus. It is noticeable that I do not always get the same results from one subcorpus to another in the same social category. However, there are a few patterns that emerge, such as the recurrence of the subject pronouns "vous", "je" and "nous" in the upper-group speech and preference in lower-group speech for "tu", "il" and "on". Articles are also much more salient in lower-group speech. Finally, in the upper-group word list, some lexical items ("monsieur", "voiture", "ami", "papa") emerge as being salient, while the lower-group speech is more normally characterised by the saliency of its grammatical words (pronouns, determiners, prepositions and conjunctions). The upper-group characters favour proper nouns ("papa", "Pierre", "Loulou"). The upper group also shows a predilection for verbs ("sera", "peut", "es", "as") rather than nouns. Particles like "alors", "mais", "plus", "puis" and "quoi", as well as the intensifiers "bien" and "très", are found to be salient in lower-group speech. "Mais" and "quoi" are salient in the upper-group speech in some films.

1.2. Comparison of Subcorpus 1 with Subcorpus 2

In the following table, I compare Subcorpus 1 with Subcorpus 2. **Column 1** presents the ten most significant words. **Column 2** gives the frequency of words in Subcorpus 1 as percentages. **Column 3** gives the frequency of the words in Subcorpus 2 as percentages. **Column 4** indicates the probability that the frequency of the word is different in the two corpora due to chance alone. The smaller the figure the more likely the frequency difference reflects a genuine dissimilarity between the two subcorpora.

Words	Frequency in Subcorpus 1	Frequency in Subcorpus 2	Probabilities
Je	2.07%	0.85%	P= 0.000
Vous	3.13%	1.74%	P= 0.000
Nous	0.5%	0.15%	P= 0.000
Oh	0.80%	0.38%	P= 0.000
De	1.93%	1.3%	P= 0.000
Très	0.21%	0.4%	P= 0.000
Mais	0.93%	0.55%	P= 0.000
Monsieur	0.35%	0.13%	P= 0.000
Euh	0.13%	less than 0.1%	P= 0.000
Oui	0.72%	0.43%	P= 0.000

Table 1

The personal pronouns “je”, “vous” and “nous” are the most salient words when my two main subcorpora are compared. The first person singular and plural is therefore more frequent in the upper-group speech. This shows that the upper group has recourse to more monologic forms than the lower-group speakers. The most notable, though predictable, finding that differentiates the lower and the upper group is the tendency for members of the latter to use negative politeness formulae with this use of “vous” and “monsieur”. “Monsieur” emerges in seventh position with 0.35% of frequency in the upper-group subcorpus. The table also shows that the upper group favours interjections of “hesitations” (“euh”) and surprise (“oh”). The saliency of the adverb “très” suggests greater involvement and could indicate that the upper-group speech is slightly more emphatic.

Conclusion:

Scott’s Wordlist has given us an idea of which words are the most salient in the film corpus. The saliency of articles does not seem at first glance to reveal anything new. My list offers a starting-point for further research on collocational patterns (Butler 1998, p.2). The word “collocation” can be used in a purely linguistic context to define “lexical patterning around the

syntagmatic axis" (Firth 1957 quoted by Butler 1998, p.1). Lexical, as well as grammatical items, can be investigated not only quantitatively but also according to their "collocational framework", that is to say the words with which they combine in the syntagm (Butler 1998, p.1). A concordance programme could carry this analysis further by investigating the phrasal structures in which these lexical items are used (Gledhill 1995, 1999).

2.0. Frequency

Frequency lists of spoken French were compiled by Guiraud (1954), the authors of *Le français élémentaire* (1964) and Muller (1967, 1968). In a frequency list of words ranked in decreasing order, Mitterand points out that "les cent premiers mots recouvrent 60% de la totalité des mots du texte dépouillé [...] les 1000 premiers mots 85%, les 4000 premiers 97.5% etc" (1963, p.15). Mitterand makes a distinction between "disponibilité" and "fréquence". By "disponibilité" is meant the words that are "probables, disponibles, usuels pour un sujet" (*ibid.*, p.13) compared to their frequency in a given speech.

The following exercise will look at the core vocabulary of Subcorpus 1 (upper-group speech) and Subcorpus 2 (lower-group speech). Table 1 presents in ranked order the first hundred most frequent lexemes in Subcorpus 1 (column 2) and in Subcorpus 2 (column 4) with the frequency of each item (columns 3 and 5). At the same time, these findings are compared with the data from the Corpus d'Orléans (Biggs & Dalwood, 1976) and with Baudot's results (1992) obtained from a contemporary corpus of written French. Baudot's corpus was compiled in 1967 in the *Bureau des langues du gouvernement du Canada* (Baudot 1992, p.9) and is made up of 803 samples of literary (rather than oral) texts (*ibid.,* p.14). The last column gives Baudot's frequency ranking.

I make no distinction between the different grammatical forms of a word. For example, the result found for "que" accumulates that of the conjunction, the relative and the pronoun. Baudot, on the other hand, separates the different functions of the word, thus introducing a certain level of disparity between the corpora compared.

Number	Upper group	*Frequency*	Lower group	*Frequency*	**Orléans**	**Baudot's corpus**
1	vous	*715*	a	*6868*	**est**	**de**
2	a	*570*	est	*1382*	**a**	**le (article)**
3	est	*504*	pas	*1012*	**et**	**être**
4	je	*475*	vous	*896*	**pas**	**un**
5	de	*446*	le	*814*	**de**	**à**
6	pas	*411*	la	*688*	**on**	**et**
7	le	*292*	de	*634*	**la**	**les**
8	que	*285*	tu	*634*	**le**	**il**
9	la	*228*	un	*612*	**oui**	**des**
10	y	*220*	que	*580*	**euh**	**que (conj.)**
11	et	*216*	il	*554*	**y**	**ne**
12	mais	*216*	on	*536*	**les**	**en**
13	ce	*197*	et	*504*	**des**	**se**
14	moi	*194*	les	*470*	**un**	**son**
15	oh	*185*	moi	*430*	**que**	**du**
16	un	*172*	je	*428*	**vous**	**au**
17	il	*171*	en	*414*	**alors**	**dans**
18	oui	*167*	ah	*362*	**mais**	**qui**
19	non	*163*	ce	*352*	**qui**	**ce**
20	ah	*161*	ai	*312*	**ouais**	**je**
21	bien	*156*	une	*292*	**je**	**pour**
22	ai	*146*	pour	*288*	**en**	**pas**
23	en	*145*	alors	*284*	**une**	**la**
24	tout	*145*	mais	*268*	**ce**	**ce**
25	une	*135*	tout	*288*	**dans**	**tout**
26	on	*130*	qui	*264*	**moi**	**plus**
27	tu	*125*	des	*254*	**pour**	**par**
28	me	*122*	non	*248*	**non**	**elle**
29	les	*118*	avec	*236*	**quoi**	**on**
30	si	*118*	bien	*218*	**tout**	**que (pron.)**
31	nous	*116*	oui	*218*	**tu**	**sur**
32	qui	*96*	comme	*212*	**puis**	**faire**
33	alors	*92*	ben	*206*	**plus**	**mais**
34	mon	*88*	me	*202*	**bien**	**nous**
35	plus	*87*	te	*202*	**si**	**le (pronoun)**
36	pour	*85*	au	*192*	**ah**	**pouvoir**
37	comme	*82*	oh	*190*	**du**	**avec**
38	monsieur	*80*	si	*190*	**heures**	**ou**

39	dans	*74*	as	*186*	**là**	**me**
40	des	*71*	eh	*184*	**quand**	**vous**
41	avec	*66*	va	*184*	**eh**	**même**
42	ma	*65*	du	*184*	**même**	**comme**
43	votre	*62*	plus	*178*	**nous**	**lui**
44	eh	*61*	elle	*174*	**ben**	**leur**
45	suis	*61*	dans	*158*	**avec**	**y**
46	du	*60*	allez	*154*	**il**	**autre**
47	avez	*55*	suis	*152*	**sont**	**mon**
48	êtes	*54*	toi	*152*	**fait**	**dire**
49	allez	*53*	es	*152*	**comme**	**en**
50	faire	*49*	lui	*150*	**deux**	**bien**
51	fait	*49*	fait	*150*	**au**	**deux**
52	rien	*49*	quoi	*146*	**elle**	**sans**
53	très	*48*	ça	*146*	**va**	**où**
54	voilà	*45*	hein	*142*	**parce**	**devoir**
55	elle	*43*	quand	*130*	**enfin**	**grand**
56	bon	*41*	dis	*124*	**n'est**	**notre**
57	être	*41*	mon	*122*	**bon**	**celui**
58	bien	*40*	veux	*116*	**hein**	**aller**
59	dire	*38*	même	*116*	**par**	**homme**
60	Jo	*38*	faire	*114*	**mon**	**aussi**
61	même	*38*	rien	*112*	**faire**	**si**
62	enfin	*37*	ils	*112*	**ans**	**quelque**
63	hein	*36*	eu	*106*	**j'ai**	**voir**
64	ami	*36*	puis	*102*	**ou**	**savoir**
65	peut	*35*	ma	*102*	**ils**	**premier**
66	chose	*35*	se	*92*	**se**	**très**
67	deux	*35*	vais	*90*	**vas**	**falloir**
68	ici	*34*	dit	*88*	**cours**	**vouloir**
69	Loulou	*34*	aime	*82*	**dire**	**encore**
70	sais	*34*	bon	*80*	**peu**	**dont**
71	se	*34*	ou	*80*	**qu'on**	**petit**
72	cette	*33*	vas	*80*	**ne**	**peu**
73	par	*33*	Marcel	*80*	**rires**	**jour**
74	quoi	*33*	deux	*78*	**leur**	**monsieur**
75	sont	*33*	donc	*74*	**voyez**	**entre**
76	veux	*33*	être	*74*	**peut**	**an**
77	ou	*31*	pourquoi	*74*	**avez**	**nouveau**
78	va	*31*	avez	*74*	**faut**	**prendre**
79	euh	*31*	faut	*72*	**Orléans**	**après**
80	Marcel	*30*	ici	*70*	**travail**	**temps**
81	merci	*30*	tiens	*70*	**lui**	**donner**
82	quand	*30*	monsieur	*68*	**ont**	**certain**
83	sur	*29*	toujours	*68*	**tous**	**non (negation)**
84	pourquoi	*29*	tous	*68*	**aussi**	**venir**
85	voulez	*27*	voir	*64*	**beaucoup**	**vie**

86	eu	*26*	à	*64*	**gros**	**moins**
87	ils	*25*	comment	*64*	**hmm**	**de**
88	savez	*25*	par	*60*	**sur**	**moi**
89	bonjour	*25*	votre	*60*	**ville**	**monde**
90	lui	*25*	peu	*60*	**voulez**	**là**
91	peu	*25*	ta	*58*	**enfants**	**seul**
92	tiens	*24*	étais	*58*	**oh**	**trouver**
93	aussi	*24*	fais	*56*	**vingt**	**les (pronoun)**
94	faut	*24*	parce	*56*	**aux**	**ainsi**
95	jamais	*24*	vrai	*56*	**cinq**	**fois**
96	mademoiselle	*24*	coup	*56*	**elles**	**quand**
97	mes	*24*	sur	*54*	**questions**	**enfant**
98	puis	*24*	dire	*54*	**être**	**toujours**
99	donc	*23*	sans	*52*	**aux**	**trois**
100	entendu	*23*	homme	*52*	**cinq**	**heure**

Table 1 Frequency lists

2.1. Comparison of the upper-group subcorpus with the lower-group subcorpus

The words present in the lower-group frequency list are mainly tool-words with articles, prepositions, adverbs, and auxiliaries. On the whole, the upper-group frequency list present a greater number of full words and proper nouns (“monsieur”, “ami”, “Jo”, “Loulou” and “Marcel”). It also shows a higher degree of formality with items such as “vous”, “je”, “moi” and “monsieur”, as compared to the more dialogic “vous”, “tu”, “ben”, “va” and “toi” of the lower-group subcorpus.

2.2. Comparison of the top 500 words in the frequency lists of Subcorpus 1 and Subcorpus 2

Taking the total lexicon of Subcorpus 1 and Subcorpus 2, I will try to assess at what point down the frequency table the upper-group speech begins to differ from the lower-group speech. Table 1 gives the percentages of words common to the frequency lists of both the upper-group subcorpus and the lower-group subcorpus.

	% common to both subcorpora
1-100	70%
100-200	33%
200-300	21%
300-400	10%
400-500	12%

Table 1

In the top hundred words, I find the invariable core vocabulary common to both groups which amounts to 70%. The shift between the lower and the upper-group speech occurs in the next hundred words. Between rank 300 and 500, the proportion of words common to both subcorpora falls to 10%.

Conclusion

For personal pronouns, the frequency list of the upper group gives preference to “vous”, “je”, “moi” and “il”. In the lower-group list, the order of frequency for pronouns is slightly different: “vous”, “tu”, “il”, “je” and “moi”. I have seen that the frequency list of the lower-group speakers was essentially composed of pronouns, articles, prepositions and connectors. Seventy percent of the top hundred words is common to both the upper and lower-group subcorpora. The two subcorpora diverge below the 300th word on the frequency table and have no more than 10% of words in common.

3.0. Comparison with “real data”

I will in this section attempt to assess whether the film corpus exists in a world of its own or whether it reflects “real usage” reasonably well. We have compared our film corpus in terms of saliency and frequency to an authentic corpus of spoken French: the corpus d’Orléans. For this project 150 Orléanais were interviewed using a questionnaire, but the collection available to us contains only twenty-five texts. These interviews cover topics of everyday life in Orléans, work

and politics. Orléans was chosen for a sociolinguistic study mainly because of its economic, political and cultural status but also because of its proximity to Paris. It should be noted of course that we are not comparing exactly like with like: our corpus dates from 1939 and the Corpus d'Orléans was compiled in the 70s, but given the absence of a control corpus from the 1930s, there was little alternative.

The data investigated in this section only represent a small part of the corpus d'Orléans published in *Les Orléanais ont la parole: Teaching Guide and Tapescript* (Biggs & Dalwood 1976). We computerised the following twenty-five transcripts which gave us a control corpus amounting to a total of 9,904 words.

	Name	Profession	Duration	words
Text 1	M. YR	skilled worker	1mn26	287
Text 2	M. DJ	ophthalmologist	1mn 04	165
Text 3	M. EX	white-collar	1mn 28	220
Text 4	Mme PF	housewife	2mn 30	547
Text 5	M. CN	priest	3mn 19	460
Text 6	M. OH	clerk	1mn 18	219
Text 7	Mme DT	clerk in post-office	1mn 41	316
Text 8	M. TM	educational adviser	2mn 57	497
Text 9	Mlle BU	white-collar	2mn 15	412
Text 10	M. YT	foreman	1mn 25	212
Text 11	Mme DT	clerk in post-office	1mn 52	405
Text 12	M. BA	butcher	3mn 42	670
Text 13	Mme UH	retired woman	1mn 25	213
Text 14	M. YR	skilled worker	2mn20	399
Text 15	Mlle QB	pediatric nurse	2mn 08	392
Text 16	M.GD	dental surgeon	3mn 12	442
Text 17	Mlle WF	home economics teacher	4mn24	801
Text 18	M. QC	chief accountant	1mn 08	181
Text 19	Mme KH	accountant	1mn 45	304
Text 20	M. LD	engineer	2mn 10	347
Text 21	M. BA	butcher	3mn 23	682
Text 22	M. HS	senior executive	3mn 09	506
Text 23	M. TM	career advisor	1mn 53	382
Text 24	Mme PF	housewife	2mn 06	390
Text 25	Four children	Primary school children	27s	438

Table 1

To produce more meaningful comparison with our film corpus, we made our control corpus larger by adding a 3,495-word interview carried out in Sarcelles by a student of Gadet

(Université de Paris X) in 1992-3. The interviewer was of Portuguese origin and the interview informal. The combining of these two subcorpora could be criticised. The two samples of naturally-occurring conversation were recorded at different times. The register of the language is not the same in each case, one subcorpus being more formal than the other. However, the comparison with our film corpus will at least give us the opportunity to compare an artificial language with naturally-occurring speech.

Our corpus of "natural" spoken French (Corpus d'Orléans and Sarcelles) amounts to a total of 13,399 words. The following table recapitulates the number of words in each corpus:

Subcorpus 1 (upper-group corpus)	19,387
Subcorpus 2 (lower-group corpus)	20,697
Subcorpus 3 (film corpus)	64,815
control corpus (Corpus Orléans and Sarcelles)	13,399

Table 2

3.1. Saliency

3.1.1. Saliency between Subcorpus 1 and the control corpus

Words	Corpus d'Orléans-Sarcelles	Frequency in upper-group subcorpus	Probabilities
Euh	1.23%	0.13%	P= 0.000
Ouaih	0.68%	0.3%	P= 0.000
Des	1.05%	0.31%	P= 0.000
On	1.42%	0.56%	P= 0.000
Les	1.27%	0.51%	P= 0.000
Là	0.32%	0.2%	P= 0.000
Oui	1.43%	0.72%	P= 0.000
C'est	2.16%	1.28%	P= 0.000
Heures	0.33%	0.05%	P= 0.000
Alors	0.93%	0.40%	P= 0.000

Table 3

The data suggests that there is a good deal of overlap between our lower and upper-group findings. The difference comes from the emergence of "on" and "c'est" as being salient in the Corpus d'Orléans-Sarcelles.

3.1.2. Saliency between Subcorpus 2 and the control corpus

Words	Corpus d'Orléans-Sarcelles	Frequency in lower-group subcorpus	Probabilities
Euh	1.23%	0.2%	P= 0.000
Ouais	0.68%	0.4%	P= 0.000
Oui	1.43%	0.43%	P= 0.000
Heures	0.33%	0.03%	P= 0.000
Nous	0.58%	0.15%	P= 0.000
Là	0.32%	0.3%	P= 0.000
Mais	1.17%	0.55%	P= 0.000
Enfin	0.34%	0.05%	P= 0.000
Est	0.31%	0.04%	P= 0.000
De	2.04%	1.3%	P= 0.000

Table 4

The interjections "euh" and "ouais", as well as "mais" and "enfin" characteristic of spontaneous speech, occur more frequently in our Corpus d'Orléans-Sarcelles than they do in the film Corpus.

3.1.3. Saliency between Subcorpus 3 and the control corpus

Words	Corpus d'Orléans-Sarcelles	Frequency in the film corpus	Probabilities
Euh	1.23%	0.06%	P= 0.000
Ouais	0.68%	0.05%	P= 0.000
Oui	1.43%	0.61%	P= 0.000
Heures	0.33%	0.05%	P= 0.000
Orléans	0.16%	-	P= 0.000
Des	1.05%	0.43%	P= 0.000
On	1.42%	0.74%	P= 0.000
An	0.24%	0.4%	P= 0.000
Travail	0.18%	0.2%	P= 0.000
Questions	0.13%	-	P= 0.000

Table 5

From this Word list emerged words like "Orléans" and "questions" which, unsurprisingly, do not occur in our film corpus. The interjections "ouaih" and "euh" stand out with 1.23% and 0.68% respectively in the Corpus d'Orléans-Sarcelles. They are features of "unplanned spontaneous" speech rather than "planned spontaneous" speech. The pronoun "on" appears to be used more frequently in our contemporary corpus than in our film corpus.

Conclusion:

The words that appear to be salient when comparing our "fabricated" film corpus with a corpus of authentic spoken French are occasionally items that are not featured in the former. Words like "questions" and "heures" are so rare in our film database that they will be computed as salient when they occur in the control corpus. We noticed that "nous" was salient when comparing Subcorpus 1 to Subcorpus 2 and this also emerges when comparing our control corpus to Subcorpus 2. When the control corpus is compared to Subcorpus 1, the impersonal form "on" emerges as significant. The results obtained by comparing Subcorpus 1 and the control corpus show a high degree of correlation with the findings obtained for the control corpus and Subcorpus 2. Subcorpora 1 and 2 together are very similar to a modern subcorpus of natural speech as far as saliency is concerned.

3.2. Frequency

3.2.1. Comparison with the corpus Orléans-Sarcelles

The frequency lists of the upper-group and lower-group subcorpora are very similar to that of the control corpus. The contrast comes from the emergence of the personal pronouns "vous", "tu" and "je" at the top of our film list, while they are less frequently used in our modern spoken corpus. The reason is that our films are interactive conversations involving several participants, while the control corpus is mainly a monologue of one speaker. Our modern corpus also feature more interjections as well as full lexical items ("heures", "cours" and "travail").

3.2.2. Comparison with Baudot's written corpus

Comparison with Baudot's findings seeks to draw out differences between a spoken and a written corpus on the one hand and between a 1930s corpus and a more recent one on the other. The distinction between our two subcorpora and Baudot's written corpus comes mainly from the use of pronouns, interjections and forms like "oui" and "non" that are most likely to occur in a

spoken corpus. The auxiliary "être" used in the infinitive has a high rate of frequency in Baudot's data, while our data show that the usage of "a" and "avoir" is more common in spoken French. The third-group verbs "pouvoir", "faire", "dire", "devoir", "voir", "savoir", "falloir" and "vouloir" are frequent lexical items in the written corpus but are absent from our oral subcorpora.

3.2.3. Comparison of the top 300 hundred words in the fr!equency lists of Subcorpus 3 with Orléans-Sarcelles /Baudot

We now look at the proportion of words common to a) the film corpus as a whole and the Orléans-Sarcelles Corpus and b) the film corpus as a whole and Baudot's written Corpus.

	% common to film corpus and Orléans-Sarcelles	% common to film corpus and Baudot
1-100	68%	44%
100-200	28%	13%
200-300	18%	11%

Table 1

The vocabulary of the film corpus correlates more highly with the Corpus Orléans-Sarcelles than with Baudot's corpus. The major differences between the film corpus and the control corpus occur after the 100th most frequent word.

Conclusion:

The frequency lists have allowed me to compare ny 1930s film corpus to "real data" through a more recent corpus of spoken French. The other point of interest was to establish similarities and differences with a modern-day written corpus. Words that are typical of spoken French like interjections and the adverbs "oui"/"non" obviously emerge. On the whole, the statistics of the first hundred most frequent words in each subcorpus indicate that the film corpus is similar to the spoken corpus Orléans-Sarcelles but correlates less well with Baudot's written corpus.

3.3.3. Correlation coefficients

The correlation coefficient tells us whether two sets of data move together: that is whether large values of one set are associated with large values of the other (positive correlation), whether small ones are associated with large values of the other (negative correlation), or whether values in both sets are unrelated (lack of correlation). Correlation coefficients are used with the present data to determine the relationships between the frequency of one set of lexical items in a given corpus and the same set of lexical items in another corpus.

To obtain a single value for both lists correlated we used the following method. In this simplified example, only the top three words in each list are considered:

Rank word	frequency
rouge	100
orange	80
jaune	60
bleu	40
vert	20

List A

Rank word	frequency
rouge	120
jaune	60
bleu	40
orange	20
blanc	15

List B

The first stage is to isolate the words which occur in the top four of both lists: e.g. red, orange, yellow, blue. The next step is to find the correlation coefficient between the rankings of the items in the two columns.

Table 1 gives the results obtained by correlating the frequency rankings of our different subcorpora. In this table, we take into account only the first hundred shared words in each subcorpus.

	Subcorpus 1	control corpus	Baudot's Corpus
Subcorpus 1 (upper group)		0.616	0.200
Subcorpus 2 (lower group)	0.672	0.680	0.343
Subcorpus 3 (film corpus)		0.648	0.272
Baudot's Corpus		0.410	

Table 1 Correlation coefficients

One can say that the film corpus as a whole correlates highly with the Orléans-Sarcelles speech but contrasts with Baudot's. The lower-group subcorpus correlates highly with the Orléans-Sarcelles speech and less well with Baudot's corpus. The upper-group subcorpus correlates less highly with the Orléans-Sarcelles speech and even less with Baudot's corpus. This is surprising since upper-group speech might be expected to be closer to writing.

Conclusion:

The comparison of the film corpus with "real data" shows that it is closer to speech than writing.

General conclusion:

My study revealed that there were differences between upper and lower-group speech at the level of saliency and frequency. Lower-group speech which shows a high number of tool-words and interjections tends to be more dialogic, whereas upper-group speech appears monologic. Comparisons with "real", though modern, data indicated that the scripted dialogues as a whole are far from being fictitious and that they are closer to natural speech than to writing.

References

Baudot J., 1992, *Fréquences d'utilisation des mots en français écrit contemporain,* Les Presses Universitaires de Montréal, Montreal.

Biggs P.P.& Dalwood M., 1976, *Les Orléanais ont la parole: Teaching Guide and Tapescript*, London, Longman (Livre de l'élève & Livre du maître).

Butler C., 1985, *Statistics in Linguistics*, Oxford, Blackwell.
" 1998, "Collocational Frameworks in Spanish" in *International Journal of Corpus Linguistics*, 3, pp.1-32.

Gledhill C., 1995, "Collocation and Genre Analysis" in *Zeitschrift für Anglistik und Amerikanistik*, 1, pp.11-36.
" 1999, "Towards a Description of English and French Phraseology" in C. Beedham, ed., 1999, *Langue and Parole in Synchronic and Diachronic Perspective*, *Proceedings of Societas Linguisticas Europea*, XXXI, Oxford, Pergamon, pp.221-37.

Gougenheim G., Rivenc P.P., Michéa R. & Sauvageot A., 1964, *Le Français fondamental*, (*1er* et *2e degré*), Institut pédagogique national, Paris, Didier.

Guiraud P., 1954, *Les Caractères statistiques du vocabulaire*, Paris, Larousse.

Mitterand H., 1963, *Les Mots français,* Collection Que Sais-Je?, n° 270 Paris, PUF.

Muller C., 1967, *Etude de statistique lexicale, le vocabulaire du théâtre de Pierre Corneille*, Paris, Larousse.
" 1968, *Initiation à la statistique linguistique*, Paris, Larousse.

Scott M.R., 1996, *WordSmith Tools*, Oxford, Oxford University Press.
" 1997, "PC Analysis of Key Words- and Key Key Words" in *System*, 25, 1, pp.1-13.